Statistics Canada Statistique Canada

Statistics Canada
Analytical Studies Branch

Research Paper Series

Canada

MORTALITY RISK DISTRITUTIONS: A LIFE TABLE ANALYSIS

by

Geoff Rowe

No. 19

Social and Economic Studies Division
Statistics Canada
Fall 1988

The analysis presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada.

MORTALITY RISK DISTRIBUTIONS:

A Life Table Analysis

Geoff Rowe
Social and Economic Studies Division
Statistics Canada

Keywords: Life table, mortality heterogeneity, empirical Bayes, nonparametric density estimation.

AMS 1980 subject classifications: Primary 92A15, secondary 62G30

ABSTRACT

This study deals with the distribution of mortality risks in Canada in the period 1975-77. The description is based on nonparametric estimates of probability densities of risks of death in five year age groups. These estimates are derived from small area, age specific mortality data after adjustment for sampling variability.

Most variability in the mortality data represented variability in risks rather than instability in estimation from small populations. The variability in risks is represented in quantile life tables. These are collections of life tables, each of which is based on risk estimates at a fixed quantile of age specific risk distributions.

Summary measures of the distribution of risks are provided by estimates of quantiles of expectation of life at birth and of median age at death. Interquartile ranges for life expectancy are estimated to be 3.2 years for males and 2.8 years for females.

1. Introduction

This study is concerned with the statistical description of variability among small area, age specific probabilities of death ${}_nq(x)$ in the age interval $[x,x+n)$, conditional on survival to age x. In most cases, n equals five. The data, which are averages of annual data over the 1975-77 time period, comprise estimates ${}_n\hat{q}(x)$ for both sexes in each of 259 geographic areas in Canada. The geographic units are census divisions (CDs) which correspond to counties or other subprovincial administrative units. The exact definition differs among provinces.

The 259 CDs represent all regions in Canada with the exception of the Yukon and Northwest Territories and the most northerly part of Quebec (Nouveau-Quebec). Although the populations of many of the CDs are large in statisticians' terms, they are generally smaller than demographers usually consider adequate for construction of life tables. Demographers often assume that populations of 100,000 or less provide unstable estimates of mortality probabilities.

The study has two features which are not common in life table analysis. First, estimates of ${}_nq(x)$ are assumed to be subject to variability which is analogous to sampling variability. Second, the observations are assumed to come from a population which has a heterogeneous distribution of mortality risks. Heterogeneity refers to differences in risk levels and

arises from differences among individuals (e.g., diet, life style, etc.).

Each CD is treated as a sample of the Canadian population which is selected without regard to mortality conditions. Consequently, each ${}_n\hat{q}(x)$ is an average taken over individuals whose mortality risks are assumed to be heterogeneous. The sampling variance of ${}_n\hat{q}(x)$ is determined by variance of unobserved risks within CDs and by CD population size. There are marked differences in the sizes of CD populations. According to the 1976 census, the population totals of CDs in the study ranged from 1500 to 2.1 million. Thus, a description of the distribution of risk involves simultaneous use of ${}_nq(x)$ estimates which are not equally reliable.

The study is directed towards more fully describing mortality conditions within national populations. Within each age interval, nonparametric estimates of probability density functions of mortality risk are calculated. These estimates provide a flexible means of assessing inequality in mortality risks (e.g., by quantile based dispersion measures).

As an outcome of density estimation, period life tables are constructed by linking the estimates of age specific risk that correspond to a given quantile. These are termed quantile life tables. This approach differs from the more conventional linkage of estimates of age specific risk corresponding to a geographically defined population. The quantile life tables are calculated for grouped ages (i.e., forming abridged quantile life

tables). Abridged quantile life tables serve to illustrate an improved description of mortality for a heterogeneous national population. For purposes other than illustration, quantile life tables by single years of age would be preferred.

The emphasis in this article is on quantile life tables. A justification for basing them on CD data requires that allowance be made for differences in reliability among ${}_n\hat{q}(x)$. That is, the life tables should reflect heterogeneity in mortality, rather than the instability of small area estimates. A further requirement is that the variability among CD means be random, so that it is valid to describe the distribution of risks by a probability density. The two sections immediately following address these aspects of the estimation problem. Section 2 describes the ${}_nq(x)$ estimator and an approximation of its variance. In section 3, it is argued that geographic grouping of the data can provide a basis for estimating the distribution of risks in the Canadian population as a whole.

Estimation of the distribution of risks is in two stages: (1) section 4 describes adjustment of observations for sampling variability by a linear empirical Bayes procedure, and (2) smooth nonparametric estimates of the probability density of risks, by age, are obtained in section 5. Section 6 provides selected illustrative results for two male age groups. Section 7 presents results derived from selected quantile life tables. Finally, concluding remarks are presented in section 8.

2. Sampling Variability

Ascribing 'sampling' variance to estimates of mortality probabilities may be controversial, if death registration is complete and census undercoverage is negligible. In that case, mortality rates might be regarded as exact values. This view makes no allowance for nonrepeating and essentially unpredictable factors influencing a population's mortality in a given time period (e.g., extremes of weather, etc.). Brillinger (1986) provides a rigorous basis for describing 'natural variability', and applies one model to the mortality of the Canadian female population, 1926-82. A variety of alternative models are presented in the discussants' comments following Brillinger's paper.

Observed values ${}_n\hat{q}(x)_j$ for the j'th CD are represented following Chiang (1972) as (using actuarial notation):

$$
\begin{aligned}
{}_n\hat{q}(x)_j &= {}_nD(x)_j / \left(\left[{}_nP(x)_j + (n - {}_na(x))\ {}_nD(x)_j \right] / n \right) \\
&= {}_nD(x)_j / K(x)_j, \qquad (1)
\end{aligned}
$$

where, ${}_nD(x)_j$ are average annual deaths in the interval 1975-77 to those aged x to x+n, ${}_nP(x)_j$ is the 1976 census population, and ${}_na(x)$ represents average years lived in the age interval by a person who died in the interval. The denominator in Eqn. 1 is an estimate of the annual population exposed to risk. In the remainder of the paper, $K(x)_j$ will be treated as if it were the true value.

The counts of deaths by age and CD were obtained from special tabulations provided by Health Division, Statistics

Canada. Population counts were obtained from the 1976 Census database. Note that population counts have been randomly rounded to multiples of five, as required for public use census tabulations.

For age groups [0,1) and [1,5), ${}_n a(x)$ was derived from abridged male and female life tables for Canada, 1975-77 (Nagnur, 1986). For all other age groups, ${}_n a(x)$ was assumed to take the value 2.5 (i.e., assuming a uniform distribution of deaths). The latter represents a first approximation. Other, more sophisticated, approximations are available (see, for example, Keyfitz and Frauenthal (1975)), but the appropriateness of these methods applied to small populations that are subject to substantial migration can not be readily evaluated. Consequently, only a first approximation is employed in this study, a strategy advocated by Hoem and Jensen (1982) and others.

Hereafter the subscript n, denoting the width of the age interval, is dropped and notation representing dependence on age is eliminated to simplify expressions. All subsequent analysis applies independently to each age interval.

A statistical model of the variability of $\hat{q}_j$ can be expressed in the following terms. Let Q_{ij} be a Bernoulli random variable representing possible mortality outcomes for the i'th individual in the j'th CD in the appropriate age interval. Q_{ij} takes the value 1 (representing a death) with probability q_{ij} and the value 0 with probability $1-q_{ij}$. Thus, q_{ij} represents a personal risk of death.

The expected count of deaths in the j'th CD is given by the sum (over i) of personal risks. The observed count (i.e., D_j) corresponds to the sum of realizations $\hat{q}_{ij}$ (taking value 1 if the person died and 0 otherwise) of the random variable Q_{ij}, so that Eqn. 1 provides an estimator of average risk. The true variance of the observed count (assuming independent risks) is given by:

$$\text{Var}(D_j) = \sum q_{ij} (1 - q_{ij}), \qquad (2a)$$

which may be rewritten as:

$$\text{Var}(D_j) = K_j \, q_{\cdot j} \, (1-q_{\cdot j}) - \sum (q_{ij} - q_{\cdot j})^2, \qquad (2b)$$

where $q_{\cdot j}$ is the average of personal risks. Eqn. 2b shows that a binomial variance of D_j gives the maximum variance among all distributions of K_j personal risks with the average $q_{\cdot j}$. The accuracy of the binomial approximation may be improved by reducing the width of age intervals, as long as K_j remains large. The result depends on all q_{ij} becoming small in the limit. Note, however, that the assumption of independent risks may be too strong.

The remainder of this paper will concern itself with subsequent stages of analysis, under the assumption that the binomial approximation is adequate. Then, the approximation to the sampling variance of $\hat{q}_j$ is $\hat{q}_j(1-\hat{q}_j)/K_j$. Note that differences among K_j (i.e., sample size) contribute to differences in reliability among $\hat{q}_j$, and that is the basis for describing the variance component as 'sampling variance'.

3. Interpreting CD Mortality Differentials

The construction of regional life tables and the presentation of mortality data in atlases can give the impression that it is appropriate to treat geographically distinct populations as if they were closed to migration or as if mortality differentials somehow depend directly on region. But, since the geographic boundaries of CDs are determined without reference to mortality conditions, differences among $\hat{q}_j$ do not represent systematic differences among CD populations.

Even where some mortality risk factors are clearly associated with geography, it does not necessarily follow that geography provides an important basis for stratifying the population at risk (i.e., forming subpopulations with relatively homogeneous risks). To the extent that the risk factors involve personal habits or life style, they can be expected to vary substantially within CD populations.

One illustration of the issue is the case of malignant cutaneous melanoma. Cumulative exposure to the sun is a risk factor in some skin cancers, and purely geographic factors such as altitude and climate may be directly relevant. However, overriding influences in the case of melanoma involve individual sensitivity and intermittency of exposure. Consequently, homogeneous risk strata could be more accurately represented in terms of occupation (indoor workers at greater risk than outdoor workers), sensitivity (fair skin, tendency to burn, etc.), and

socioeconomic status (recreational tanning, winter vacations, etc.), than solely in terms of geography (Elwood et al. (1985)).

The characteristics of individuals, whose membership in one CD population rather than another is the outcome of arbitrary grouping, are a source of non-systematic variability among estimates of average risk. The grouping may correspond to a random partition of the Canadian population. Then, the distribution of CD averages will reflect the distribution of risks of typical individuals in the national population. That is, a probability density is an appropriate description of the distribution of CD averages, since the source of variability has a frequency interpretation.

4. Linear Empirical Bayes Adjustment for Sampling Variability

Linear empirical Bayes procedures (LEB) may be used to estimate vectors of random means. A simple form of LEB procedure can be represented as estimation of a random parameter vector $\emptyset$ by regression on a vector of observed data y (Robbins, 1983). The goal is to minimize $E(\emptyset(y)-\emptyset)^2$, where $\emptyset(y)$ is a linear function of y estimating $\emptyset$. In the present application, LEB procedures involve shrinking each $\hat{q}_j$ toward the centre of the distribution. The magnitude of shrinkage is large, if the sampling variance of a given $\hat{q}_j$ is large (i.e., providing an adjustment for sampling variability).

Shrinking estimates in the direction of their overall average provides a reasonable compromise between using the overall average to represent every CD (i.e., effectively assuming

that variance between CDs is due to sampling variability) or accepting each CD estimate without modification (i.e., effectively assuming that CDs represent unrelated populations).

A variance stabilizing transformation is employed so that each $\hat{q}_j$ will have (asymptotic) sampling variance independent of $q_{\cdot j}$. Variance stabilization is given by Anscombe's (1949) modification of the arcsin transformation:

$$\hat{t}_j = 2 \arcsin\left(\sqrt{(D_j + 3/8)/(K_j + 3/4)}\right). \quad (3)$$

Anscombe's modification gives added stability for CDs with small K, but its effect is negligible for most CDs in this application. Similar use of an arcsin transformation applied to binomial data in an LEB context is found in Efron and Morris (1975).

Variability among CD averages is represented in the model:

$$\hat{t}_j = t_{\cdot} + (t_j - t_{\cdot}) + u_j = t_{\cdot} + e_j, \quad (4)$$

where, $t_j = 2 \arcsin\left(\sqrt{q_{\cdot j}}\right)$.

Note, $\mathrm{Var}(e_j) = v + 1/K_j$, where $1/K_j$ represents binomial (sampling) variance of u_j given the arcsin transformation and v represents variance in excess of sampling variance (i.e., variance among CD deviates $(t_j - t_{\cdot})$). Further note that $\mathrm{Cov}(\hat{t}_j, t_j) = v$.

Estimates of $t_{\cdot}$ and v are obtained iteratively. The estimate $\hat{t}_{\cdot}$ employs the inverse of the estimated variance of each $\hat{t}_j$ as a weight:

$$\hat{t}_{\cdot} = \sum (\hat{t}_j \, wm_j) \, / \sum wm_j, \quad (5)$$

where $wm_j = 1/(\hat{v} + 1/K_j)$ with summation over m CDs. Note that if $\hat{v}$ were large enough the weights would be nearly equal (i.e., approximately independent of K_j).

An estimate of v may take the same form as would a maximum likelihood estimate assuming normality with correction factor m/(m-1). A convenient computational form is:

$$\hat{v}_{i+1} = [m/(m-1)] \lesssim [(\hat{t}_j-\hat{t}) wv_j]^2 / \lesssim wv_j , \qquad (6)$$

where $wv_j = [1/(1 + 1/(\hat{v}_i K_j))]$, with $\hat{v}_{i+1}$ a current estimate and $\hat{v}_i$ the previous estimate. Note, $E(\hat{v}) = v$ whether the deviates $(\hat{t}_j-\hat{t})$ are normal or otherwise. Given a starting value for $\hat{t}$., Eqn. 6 may be solved iteratively. Then $\hat{t}$ is estimated from Eqn. 5 with new weights, and the cycle is repeated to convergence.

Table 1 presents selected results of the variance computations. The table shows coefficients of variation (CV) and percentages of variance in excess of sampling variance (i.e., termed % Excess) for each age group. These are summary measures which provide a comparison of standard non-sampling deviations to the average and excess variance relative to sampling variance, respectively. CVs range between 8% and 28%, indicating that deviations from the average are typically substantial relative to the average. With the exception of ages 0-1, the % Excess values are all greater than 70% (i.e., indicating that most of the variability exhibited in the data is non-sampling variability). These results clearly indicate that there are important differences among the mortality levels exhibited in the CD data.

Correspondingly, the overall average is not a uniformly adequate indicator of risks experienced in the population.

Standard LEB estimates of risk (q') may be obtained from:

$$t'_j = 2 \arcsin(\sqrt{q'_j}) = \hat{t} + [\hat{v} / (\hat{v} + 1/K_j)]\, e_j, \qquad (7)$$

where the shrinkage factor applied to e_j estimates the ratio of $\mathrm{Cov}(\hat{t}_j, t_j)$ to $\mathrm{Var}(\hat{t}_j)$ (i.e., analogous to a regression slope). Extending the regression analogy, residual variance $\mathrm{Var}(t'_j - t_j)$ may be expressed as $\mathrm{Var}(t_j) - \mathrm{Cov}(\hat{t}_j, t_j)^2/\mathrm{Var}(t_j)$ and is estimated by $[\hat{v} / (\hat{v} + 1/K_j)]\, (1/K_j)$.

A number of authors (e.g., Louis (1984) and Gaver and O'Muircheartaigh (1987)) have expressed concern that standard LEB estimators shrink too far. For example, the observation that the sample variance $\sum (t'_j - \hat{t})^2/m-1$ is less than $\hat{v}$ implies that the empirical distribution function (EDF) of t'_j will not approximate the EDF of t_j. Louis has proposed a restricted LEB estimator which provides an approximation of the EDF of t_j that minimizes distance from the standard LEB estimates (i.e., minimizing $\sum wc_j (t^*_j - t'_j)^2$):

$$t^*_j = 2 \arcsin(\sqrt{q^*_j}) = [wc_j/(wc_j+d_1)]\, t'_j + [d_2/(wc_j+d_1)], \qquad (8)$$

where wc_j is an estimate of $1/\mathrm{Var}(t'_j - t_j)$.

The parameters d_1 and d_2 are lagrangian multipliers chosen so that the sample average $t^*_\cdot = \sum t^*_j/m$ and the sample variance $\sum (t^*_j - t^*_\cdot)^2/m-1$ match $\hat{t}$ and $\hat{v}$, respectively. Values of d_1 for each sex and age group were obtained by a secant algorithm. At convergence approximate values of d_1 gave sample

standard deviations that differed from $\hat{v}$ by less than 0.00001 (i.e., less than 0.01% error in all cases). Parameter d_2 given d_1 was calculated by rearrangement of Eqn. 8 and provided sample averages that agreed with $\hat{t}$ exactly in each case.

The restricted LEB estimates t^*_j are reasonable estimates of risk quantiles, in as much as they are each close to the corresponding standard LEB estimate and agree closely with the estimated mean and variance of $\hat{t}_j$. Thus, the t^*_j provide a basis for approximating probability density functions of mortality risk given adjustment for differences in the reliability of $\hat{q}_j$.

5. Risk Density Function Estimates

Non-parametric estimation of a probability density function of mortality risk is intended to provide an indication of the characteristics of the distributions that would have to be accounted for in an appropriate parametric model. In that sense, this stage of analysis is exploratory. However, density estimation can also provide pseudo-likelihood location and scale estimates of mortality risk, and aid interpolation to equally spaced risk quantiles.

The goodness of fit of a simple parametric reference model of t* may be assessed from the linearity or lack of linearity in the association between t^* order statistics and expected reference quantiles. In the present application, the reference quantiles (denoted hereafter as z) are from a beta density function. The beta density has the form:

$$f(z) = [\ z^{b-1}\ (1-z)^{c-1}\ ]\ /\ \mathrm{Beta}(b,c), \qquad (9)$$

where Beta(b,c) is the beta function. Beta densities can accommodate a wide variety of distributional shapes and are suitable for modeling bounded random variables (t^* is bounded by zero and π). The parameters b and c are chosen to match $\hat{t}/\pi$ and $\hat{v}/\pi^1$.

Silverman (1985) describes a smoothing cubic spline algorithm for nonparametric regression which will be used here to provide both smooth estimates of t^* and smooth estimates of first derivatives dt^*/dz. The algorithm minimizes the vector function:

$$(1-h)\ [t^*-t^*(z)]^T\ W\ [t^*-t^*(z)]\ +\ h \int t^{*\prime\prime}(z)^2\ dt^*(z), \qquad (10)$$

where t^* is a vector of order statistics, $t^*(z)$ is a vector of fitted values (from regression on z), $t^{*\prime\prime}(z)$ represents second derivatives of the fitted values, and W is a diagonal matrix of weights. The derivatives $dt^*(z)/dz$ estimate the ratio of the (known) probability of z to the probability of $t^*(z)$. Thus, nonlinearities in the association between t^* and z can be represented by an empirical density differing from the fitted beta density.

The parameter h in Eqn. 10 is a smoothing parameter that differentially weights local accuracy and global smoothness. As h approaches 1.0, the algorithm represents the relation between t^* and z as a location-scale transformation (i.e., linearity implies $t^{*\prime\prime}(z)=0$). As h approaches 0.0, the algorithm implicitly includes as many parameters as there are observations (i.e., providing an interpolation where $t^*-t^*(z)=0$). An

appropriate value of h can be obtained by minimizing cross-validation errors (Silverman, 1984).

As the fitted values are estimated quantiles, W is available from the inverse of estimates of $t^*(z)$ variances:

$$\mathrm{Var}(t^*(z)) = s^2 \left[(p(1-p)) / (m\, f(t^*(z))^2) \right], \qquad (11)$$

where p is an appropriate EDF estimate, m is the number of observations, and s^2 is a scalar (Parzen, 1979). The probability ordinates $f(t^*(z))$ are obtained from the probability ordinates of z divided by the corresponding estimate of $dt^*(z)/dz$. Thus, the spline smoothing algorithm can be iterated, updating W each time from improved estimates of $f(t^*(z))$ and s^2. Minimization of cross-validation errors and iterative use of Eqn. 11 led to values of h (for each age group) that were in the range 0.76 to 0.99.

Details of similar use of order statistics and model quantiles for nonparametric density estimation are found in Bofinger (1975) and Parzen (1979). The use of spline smoothing is justified by the interpretation of the spline as a moving average (kernel) smoother with a local bandwidth proportionate to the local variance (Silverman, 1984).

The estimates $t^*(z)$ are subsequently employed to construct life tables. For these purposes, $\bar{q}^*$ is obtained by inverting the arcsin transformation (i.e., $\bar{q}^* = [\sin(t^*(z)/2)]^2$) and estimates of the probability density function of $\bar{q}^*$ are obtained by applying the Jacobian of the inverse transform to $f(t^*(z))$.

6. Illustrative Results

In this section, intermediate results are presented which illustrate the analysis. Figure 1 provides a plot of $\hat{q}$ versus $\tilde{q}^*$ for males in age groups 0-1 and 50-54. These age groups were chosen because the distributions were well separated, but not so much as to completely obscure detail (i.e., by reducing the resolution of the plot). Corresponding plots of empirical density ratios are presented in Figure 2. The density ratios, estimated from the smooth first derivatives obtained above, emphasize differences that may exist between empirical mortality risk density functions and transformations of the fitted beta density functions. In that sense, the density ratios aid an evaluation of goodness of fit. Note that the plotted density ratios have been adjusted for the constant scale factor π.

A general index of the effects of shrinking and smoothing can be provided by the slope of a resistant line fitting $\hat{q}_j$ as a function of $\tilde{q}^*_j$. The slope for males aged 0-1 is 1.33 which indicates that the dispersion of unadjusted mortality risks is typically 33% greater than the adjusted. The corresponding slope for males aged 50-54 is also 1.08. The maximum slope for males is 1.33 (ages 0-1), while the maximum for females is 1.21 (ages 35-39). The corresponding minimum values are 1.03 for males aged 20-24, and 1.06 for females aged 65-69. The effect of the adjustments on the dispersion of mortality risk estimates is not large, an observation which is not consistent

with the assumption that small area mortality estimates are generally unstable.

Figure 1 illustrates the effects of shrinking and smoothing on the $\hat{q}_j$. In age group 0-1, the upper tail (including seven or more possible outliers) has been drawn in substantially. Note that this age group has the smallest proportion of non-sampling variance of all age groups, as indicated by the % Excess values in Table 1. Consequently, this age group is more strongly influenced by shrinking than other age groups, as indicated by the resistent line slope above.

Figure 2 provides an indication of the qualitative differences there might be between the true distribution of mortality risks and the beta density that was chosen as a reference distribution. For example, density ratios for males aged 50-54 at generally close to 1.0. Thus, for this age group, the beta density may provide an adequate probability model. Perhaps more to the point, a two parameter (i.e., mean and variance) model appears adequate to describe the distribution of mortality risks for this age group. By contrast, density ratios for ages 0-1 suggest mortality risks among infants that are less skewed and have shorter tales than would be represented in a beta density. This is implied by the shape of the density ratio function, that is by the peak in the centre (values > 1.0) and low values (values < 1.0) in the tails. Density ratios similar to males aged 0-1 (i.e., indicating a lack of fit) are more common among other age groups than are density ratios similar to males

age 50-54. This would imply that three or four parameters may be required to approximate the distribution of mortality risks for most age groups.

The results presented here are sensitive to the shrinkage parameters in Eqn. 8 (i.e., d_1 and d_2) and to the choice of the smoothing parameter in Eqn. 10 (i.e., h). Thus, while the results may suggest that a four parameter model would be more appropriate tha a two parameter model, they do not provide a basis for significance tests (e.g., likelihood ratio tests).

7. Quantile Life Tables

Period life tables represent mortality conditions in terms of completed synthetic life times (e.g., forming survival curves from the product integral of age specific survival probabilities $(1-\hat{q})$). Quantile life tables are a generalization of period life tables constructed on sections through the smoothed age specific mortality distributions of $\tilde{q}^*$. These sections link corresponding quantiles of different age groups. The resulting survival curves can be used to assess variability in duration of life (i.e., life expectancy at birth, or the median age at death).

Table 2 provides a comparison of previously published mortality probabilities (Nagnur, 1986 - termed AGGREGATE estimates in Table 2) with estimates employing the smoothed CD mortality probabilities and the empirical density functions. This comparison is intended to validate the life table results

presented here by demonstrating that the estimates are comparable in magnitude to more conventional estimates. Moreover, for most age groups, the order of MEAN, MEDIAN, and MODAL values clearly indicates positive skewness in the risk distributions.

However, none of the estimates in Table 2 are strictly comparable as estimates of the same measure of location. In particular, the AGGREGATE and MEAN estimates differ in as much as the former may be represented as a weighted average of $\hat{q}_j$ with K_j as weights and the latter is related to a weighted average with the reciprocals of empirical variances as weights. A weighting scheme similar to that employed in Eqn. 5 could result in markedly smaller differentials among weights compared with differentials among population sizes. Thus, the observation that estimates employing CD data are frequently higher than the more conventional AGGREGATE estimates might imply a bias in the AGGREGATE estimates. Because of the differences in these particular estimation procedures, the issue of which estimator may have the greater bias is too complex to resolve here. Nevertheless, weighting data by population size may produce inappropriate results.

Table 3 provides estimates of average and median duration of life from selected quantile life tables. These suggest that the variability in risks experienced within the Canadian population is consistent with a wide range of durations of life. Partial verification of these results is provided by comparing the low tail estimates to a group within the Canadian

population that is known to have had high mortality in the same time period. The expectation of life in the Registered Indian population in 1976 (representing about 1% of the Canadian population) was 60 years for males and 66 years for females (Rowe and Norris, 1985). Estimates based on quantile life tables are less extreme than the estimates for the Registered Indian population, which may indicate that the quantile estimates are conservative.

Table 3 provides a perspective on recent Canadian mortality trends. Over the past two decades, life expectancy in Canada has been increasing by at least one year over each five year time interval. Such increases are not large in comparison to the quantile ranges in Table 3. As such, it is worth asking whether these trends result from general improvements in health or from improvements that are specific to a part of the population. That is, would risk distributions from earlier time periods be shifted upwards or would their upper tails be stretched relative to the 1976 distributions? The distinction has bearing on our expectations for future trends in mortality.

Quantile life tables serve to focus attention on two elements of the description of mortality conditions in heterogeneous populations. That is, on the need to allow for differences among individuals of the same age, and for differences that might exist in the same group of individuals at different ages. Survival curves may be constructed by specifying both a distribution of risk at each age and specifying a path

through the succession of risk distributions. Presumably, smokers who stop smoking follow a different path than smokers who do not. Quantile life tables represent paths through the risk distributions that are in a sense determined at birth.

8. Discussion

The key findings of this study might be summarized as follows:

(a) Most of the variability among small area, age specific mortality probabilities reflects variability in mortality risks, rather than instability in small area estimates.

(b) Differences among CD averages are large enough to be of importance for descriptive or policy purposes, even after adjustment for sampling variance and after smoothing. For example, recent national mortality trends represented by changes in life expectancy might reflect contraction in the tails of the risk distributions, rather than shifts at the centre.

(c) Extreme quantile life tables provide a direct description of the range of mortality risk in a heterogeneous population by representing the best and worst synthetic life times that are consistent with the data. Quantile life tables also provide estimates of dispersion near the centre of mortality risk distributions.

The interpretation of these results depends on the assumption that CD populations could be representative samples of individuals with independent risks. This view is very different from one which explicitly or implicitly assumes region specific

effects. However, it is also a view which more nearly accords with mortality conditions in which contagious disease plays a relatively minor role while life style plays a major one.

These results might provide encouragement for further analysis of finely partitioned mortality data. Extensions might make use of partioning variables that can not be viewed as non-systematic (e.g., marital status), in addition to use of finer geographic resolution. But, for the present, replication of this work over a number of time periods might be the most promising extension.

ACKNOWLEDGEMENTS

I thank my colleagues Jane F. Gentleman and David Binder, whose comments contributed to improvement of both the technical content and presentation of this paper. Thanks are also due to two anonymous referees for their comments.

References

Anscombe, F.J. (1948), "The transformation of poisson, binomial, and negative binomial data" Biometrika 35, 246-54.

Bofinger,E. (1975), "Estimation of a density function using order statistics" Austral.J. Statistics 17, 1-7.

Brillinger, David R. (1986), "The Natural Variability of Vital Rates and Associated Statistics" Biometrics 42, 693-734.

Chiang, C.L. (1972), "On constructing current life tables" JASA 67, 538-541.

Efron, Bradley and Morris, Carl (1975), "Data Analysis Using Stein's Estimator and Its Generalizations" JASA 70, 311-319.

Elwood,J.M., Gallagher,R.P., Hill,G.B., and Pearson, J.C.G., (1985), "Cutaneous Melanoma in Relation to Intermittent and Constant Sun Exposure - The Western Canada Melanoma Study" Int.J. Cancer 35, 427-433.

Gaver,Donald P., and O'Muircheartaigh, I.G. (1987), "Robust Empirical Bayes Analyses of Event Rates" Technometrics 29, 1-15.

Hoem, Jan M. and Jensen, Ulla Funck (1982), "Multistate Life Table Methodology: A Probabilist Critique". from Multidimensional Mathematical Demography, Kenneth C. Land and Andrei Rogers editors, Academic Press, New York.

Keyfitz, Nathan and Frauenthal, James (1975), "An Improved Life Table Method" Biometrics 31, 889-899.

Louis, Thomas A. (1984), "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods" JASA 79, 393-398.

Nagnur, Dhruva (1986), Longevity And Historical Life Tables: 1921-81 (Abridged) Canada and the Provinces, Catalogue 89-506, Statistics Canada, Ottawa.

Parzen, Emanuel, (1979), "Nonparametric Statistical Data Modeling" JASA 74, 105-131.

Robbins, Herbert, (1983), "Some Thoughts on Empirical Bayes Estimation" Annals of Statistics 11, 713-723.

Rowe, Geoff and Norris, M.J., (1985), "Mortality Projections of Registered Indians, 1982 to 1996" project report for Research Branch, Corporate Policy, Indian and Northern Affairs Canada, Ottawa.

Silverman, B.W. (1984), "Spline Smoothing: The Equivalent Variable Kernel Method" Annals of Statistics 12, 898-916.

Silverman, B.W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" J.R. Statis. Soc. B, 47, 1-52.

TABLE 1

Summary Measures of Variability
in Census Division Mortality Risk
(Arcsin Transforms t_j of q_j)

AGE GROUP	MALES Coefficient of Variation	MALES % Excess	FEMALES Coefficient of Variation	FEMALES % Excess
0-1	14.7	50.1	16.3	50.1
1-4	28.4	82.1	26.2	74.6
5-9	24.8	77.8	27.5	76.1
10-14	24.7	79.7	27.3	76.0
15-19	23.0	91.3	25.0	81.5
20-24	25.4	92.8	27.3	81.0
25-29	21.4	86.7	26.0	78.7
30-34	20.5	84.0	24.0	76.9
35-39	19.1	82.7	20.2	73.6
40-44	16.4	82.7	16.8	72.3
45-49	13.7	83.5	17.6	80.7
50-54	11.5	84.2	14.3	78.5
55-59	11.3	87.6	13.2	81.9
60-64	9.6	87.8	12.1	82.7
65-69	8.7	86.7	12.5	84.9
70-74	9.3	89.5	9.7	80.6
75-79	9.0	88.2	10.1	75.7
80-84	8.2	84.6	11.1	88.0
85-89	9.2	85.0	9.1	77.7

Coefficient of variation $= 100\sqrt{\hat{v}} / \hat{t}$

% Excess - Variance in Excess of Sampling Variance as a Percent of Total Variance $= 100\ (m\ v) / \sum (v + 1/K_j)$

$\hat{t}$ - estimated average, $\hat{v}$ - Non-sampling variance,
K_j - population size, m - 259 Census Divisions

Table 2

Comparison among Location Estimates
for Mortality Probabilities q

AGE	AGGREGATE	MEAN	MEDIAN	MODAL
MALES				
0-1	0.0141	0.0149	0.0146	0.0143
1-4	0.0032	0.0042	0.0039	0.0034
5-9	0.0023	0.0033	0.0032	0.0029
10-14	0.0021	0.0030	0.0028	0.0025
15-19	0.0075	0.0092	0.0087	0.0077
20-24	0.0093	0.0117	0.0111	0.0099
25-29	0.0074	0.0092	0.0088	0.0081
30-34	0.0078	0.0098	0.0095	0.0089
35-39	0.0107	0.0123	0.0118	0.0110
40-44	0.0164	0.0183	0.0178	0.0169
45-49	0.0273	0.0284	0.0279	0.0270
50-54	0.0439	0.0442	0.0437	0.0428
55-59	0.0693	0.0694	0.0687	0.0675
60-64	0.1053	0.1050	0.1054	0.1068
65-69	0.1557	0.1507	0.1499	0.1486
70-74	0.2281	0.2224	0.2219	0.2224
75-79	0.3234	0.3179	0.3165	0.3174
80-84	0.4508	0.4467	0.4453	0.4459
85-89	0.5989	0.6214	0.6207	0.6131
FEMALES				
0-1	0.0114	0.0123	0.0120	0.0115
1-4	0.0025	0.0035	0.0033	0.0032
5-9	0.0015	0.0024	0.0023	0.0020
10-14	0.0013	0.0020	0.0018	0.0016
15-19	0.0026	0.0036	0.0034	0.0030
20-24	0.0027	0.0036	0.0034	0.0030
25-29	0.0028	0.0036	0.0034	0.0028
30-34	0.0038	0.0047	0.0044	0.0039
35-39	0.0059	0.0072	0.0070	0.0067
40-44	0.0091	0.0103	0.0101	0.0097
45-49	0.0152	0.0151	0.0147	0.0144
50-54	0.0218	0.0227	0.0225	0.0220
55-59	0.0339	0.0342	0.0338	0.0330
60-64	0.0512	0.0513	0.0511	0.0515
65-69	0.0800	0.0795	0.0794	0.0804
70-74	0.1267	0.1282	0.1279	0.1287
75-79	0.2010	0.2037	0.2024	0.2010
80-84	0.3222	0.3318	0.3324	0.3337
85-89	0.4866	0.4965	0.4919	0.4817

AGGREGATE - as published in Nagnur (1986) - based on the ratios of total deaths to total population.
MEAN, MEDIAN, MODAL - as determined by empirical Bayes estimates of mortality risk and empirical probability densities.

TABLE 3

Quantile Estimates of Duration of Life

Quantile	Life Expectancy at Birth		Median Age at Death	
	MALES	FEMALES	MALES	FEMALES
1%	63.4	71.4	68.9	77.0
5%	65.5	73.4	70.5	78.4
10%	66.5	74.3	71.2	79.1
25%	68.2	75.7	72.5	80.2
50%	69.8	77.1	73.8	81.1
75%	71.4	78.5	75.2	82.1
90%	72.8	79.7	76.4	83.0
95%	73.7	80.5	77.1	83.6
99%	75.3	82.0	78.4	84.8
MEAN	69.6	77.0	73.7	81.0
MODAL	70.2	77.4	74.0	81.2
AGGREGATE	70.26	77.70	73.8	81.3

MEAN, MODAL estimates obtained from mortality probabilities in Table 2.

AGGREGATE - life expectancy as published in Nagnur (1986)- median ages at death interpolated from the published survivor curve.

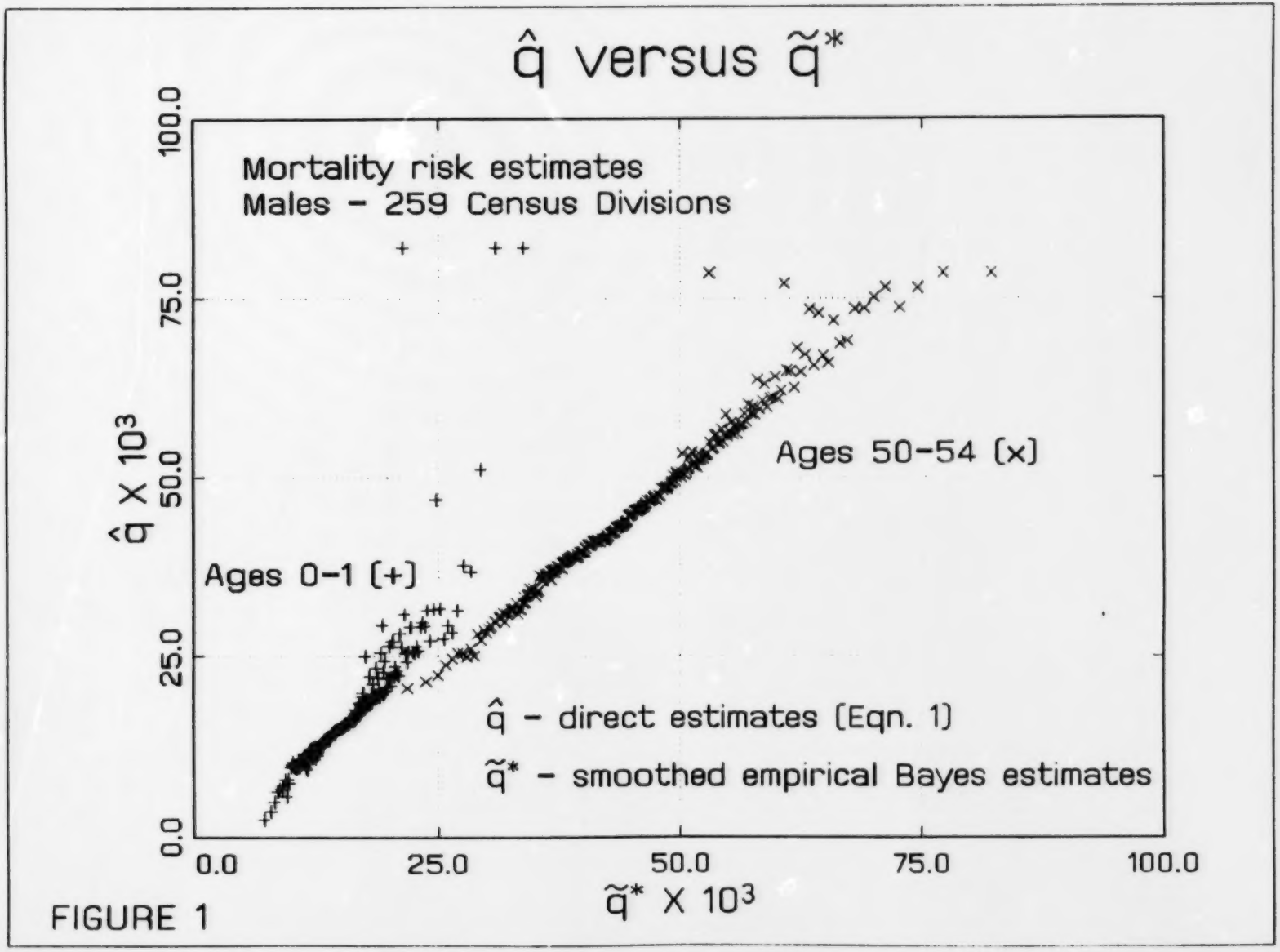

$\hat{q}$ versus $\tilde{q}^*$
Mortality risk estimates
Males - 259 Census Divisions
Ages 50-54 (x)
Ages 0-1 (+)
$\hat{q}$ - direct estimates (Eqn. 1)
$\tilde{q}^*$ - smoothed empirical Bayes estimates
$\hat{q}$ X 10^3
$\tilde{q}^*$ X 10^3
0.0
25.0
50.0
75.0
100.0

FIGURE 1

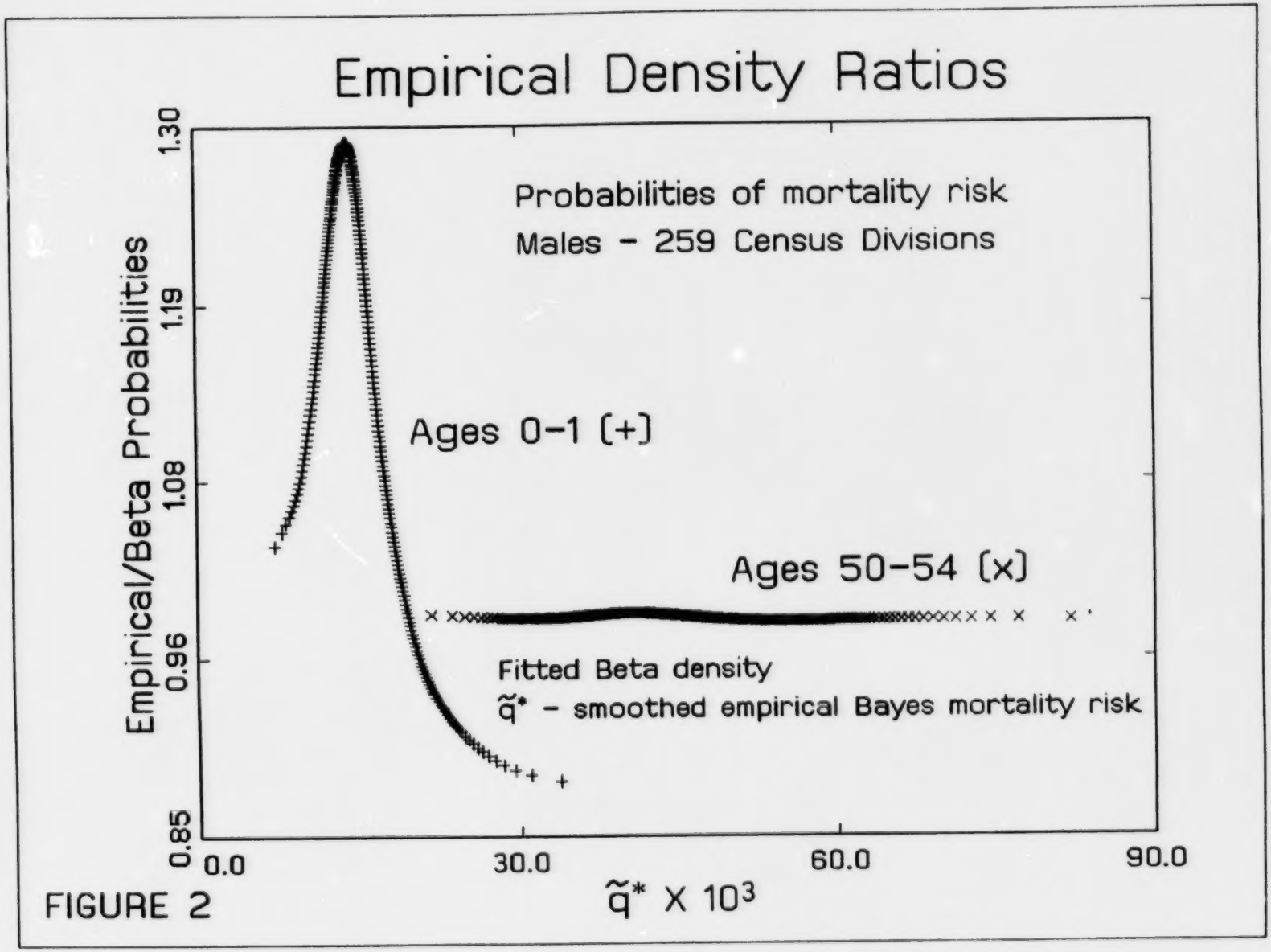

Empirical Density Ratios
Probabilities of mortality risk
Males - 259 Census Divisions
Ages 0-1 (+)
Ages 50-54 (x)
Fitted Beta density
q̃* - smoothed empirical Bayes mortality risk
Empirical/Beta Probabilities
1.30
1.19
1.08
0.96
0.85
0.0
30.0
60.0
90.0
q̃* X 10³

FIGURE 2

ANALYTICAL STUDIES BRANCH

RESEARCH PAPER SERIES

No.

1. BEHAVIOURAL RESPONSE IN THE CONTEXT OF SOCIO-ECONOMIC MICROANALYTIC SIMULATION by Lars Osberg

2. UNEMPLOYMENT AND TRAINING by Garnett Picot

3. HOMEMAKER PENSIONS AND LIFETIME REDISTRIBUTION by Michael Wolfson

4. MODELLING THE LIFETIME EMPLOYMENT PATTERNS OF CANADIANS by Garnett Picot

5. JOB LOSS AND LABOUR MARKET ADJUSTMENT IN THE CANADIAN ECONOMY by Garnett Picot and Ted Wannell

6. A SYSTEM OF HEALTH STATISTICS: TOWARD A NEW CONCEPTUAL FRAMEWORK FOR INTEGRATING HEALTH DATA by Michael Wolfson

7. A PROTOTYPE MICRO-MACRO LINK FOR THE CANADIAN HOUSEHOLD SECTOR SESSION 3, MACRO/MICRO LINKAGES - HOUSEHOLDS by Hans Adler and Michael Wolfson

8. NOTES ON CORPORATE CONCENTRATION AND CANADA'S INCOME TAX by Michael Wolfson

9. THE EXPANDING MIDDLE: SOME CANADIAN EVIDENCE ON THE DESKILLING DEBATE by John Myles

10. THE RISE OF THE CONGLOMERATE ECONOMY by Jorge Niosi

11. ENERGY ANALYSIS OF CANADIAN EXTERNAL TRADE: 1971 and 1976 by K.E. Hamilton

12. NET AND GROSS RATES OF LAND CONCENTRATION by Ray Bollman and Philip Ehrensaft

13. CAUSE-DELETED LIFE TABLES FOR CANADA (1921 to 1981): AN APPROACH TOWARDS ANALYSING EPIDEMIOLOGIC TRANSITION by Drhuva Nagnur and Michael Nagrodski

14. THE DISTRIBUTION OF THE FREQUENCY OF OCCURRENCE OF NUCLEOTIDE SUBSEQUENCES BASED ON THEIR OVERLAP CAPABILITY by Jane F. Gentleman and Ronald C. Mullin

15. IMMIGRATION AND THE ETHNOLINGUISTIC CHARACTER OF CANADA AND QUEBEC by Réjean Lachapelle

16. INTEGRATION OF CANADIAN FARM AND OFF-FARM MARKETS AND THE OFF-FARM WORK OF WOMEN, MEN AND CHILDREN by Ray Bollman and Pamela Smith

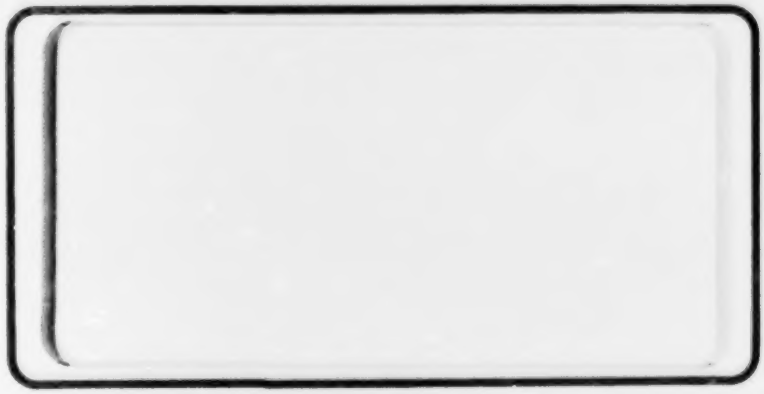
17. WAGES AND JOBS IN THE 1980s: CHANGING YOUTH WAGES AND THE DECLINING MIDDLE by J. Myles, G. Picot and T. Wannell

18. A PROFILE OF FARMERS WITH COMPUTERS by Ray D. Bollman

19. MORTALITY RISK DISTRIBUTIONS: A LIFE TABLE ANALYSIS by Geoff Rowe

For further information, contact the Chairperson, Publication Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario K1A 0T6.

Statistique Canada Statistics Canada



Statistique Canada
Direction des études analytiques

Documents de recherche

Canada

RÉPARTITIONS DES RISQUES DE MORTALITÉ:
Une analyse de tables de mortalité

par

Geoff Rowe

n° 19

Division des études sociales et économiques
Statistique Canada
automne 1988

L'auteur assume seul la responsabilité des opinions formulées dans le présent document qui ne représentent pas nécessairement le point de vue de Statistique Canada.

RÉPARTITIONS DES RISQUES DE MORTALITÉ:

Une analyse de tables de mortalité

Geoff Rowe

Division des études sociales et économiques

Statistique Canada

Mots clés: Table de mortalité, hétérogénéité de la mortalité, méthode empirique de Bayes, estimation non paramétrique de densité.

AMS 1980: classifications par matière: primaire 92A15, secondaire 62G30

RÉSUMÉ

La présente étude examine la répartition des risques de mortalité au Canada pendant la période 1975-1977. La description repose sur des estimations non paramétriques des densités de probabilité des risques de mortalité dans des groupes d'âges quinquennaux. Ces estimations sont tirées de données sur la mortalité par âge dans des petites régions, après ajustement destiné à tenir compte de la variabilité d'échantillonnage.

La variabilité dans les données sur la mortalité représentait en majeure partie la variabilité existant dans les risques plutôt que l'instabilité des estimations en provenance de populations à faible effectif. La variabilité des risques est indiquée dans des tables de mortalité quantiles, lesquelles regroupent des tables de mortalité, dont chacune repose sur des estimations des risques à des quantiles fixes des répartitions des risques par âge.

Les estimations des quantiles d'espérance de vie à la naissance et d'âge médian au décès donnent des mesures sommaires de la répartition des risques. Selon les estimations, les intervalles interquartiles pour l'espérance de vie s'établissent à 3,2 ans chez les hommes et à 2,8 chez les femmes.

1. Introduction

La présente étude examine la description statistique de la variabilité, entre de petites régions, des quotients de mortalité par âge ${}_nq(x)$ dans l'intervalle d'âge $[x, x+n)$, en cas de survie jusqu'à l'âge x. Dans la plupart des cas, n est égal à cinq. Les données, qui sont des moyennes des statistiques annuelles portant sur la période 1975-1977, comprennent des estimations ${}_n\hat{q}(x)$ pour les deux sexes dans chacune des 259 régions géographiques du Canada. Les unités géographiques sont les divisions de recensement (DR), qui correspondent aux comtés ou à d'autres unités administratives infraprovinciales. La définition exacte diffère d'une province à l'autre.

Les 259 DR représentent toutes les régions du Canada, à l'exception du Yukon, des Territoires du Nord-Ouest et de la partie la plus septentrionale du Québec (Nouveau-Québec). Bien que les populations de nombreuses DR soient importantes du point de vue du statisticien, elles sont généralement inférieures à la taille que les démographes jugent d'habitude requise pour la construction de tables de mortalité. Les démographes admettent souvent en postulat que des populations de 100 000 habitants ou moins fournissent des estimations instables des quotients de mortalité.

L'étude présente deux caractéristiques que l'on ne rencontre pas souvent dans l'analyse des tables de mortalité. Premièrement, on suppose que les estimations de ${}_nq(x)$ sont soumises à une variabilité qui est analogue à la variabilité d'échantillonnage. Deuxièmement, les observations sont censées venir d'une population qui possède une répartition hétérogène des risques de mortalité. L'hétérogénéité se rapporte aux divergences dans les niveaux de risque et provient des différences qui existent entre les individus (par exemple, régime alimentaire, mode de vie, etc.).

Chaque DR est traitée comme un échantillon de la population canadienne, qui est choisi sans tenir compte des conditions de mortalité. Par conséquent, chaque ${}_n\hat{q}(x)$ est une moyenne calculée pour des personnes dont les risques de mortalité sont censés être hétérogènes. La variance d'échantillonnage de ${}_n\hat{q}(x)$ est déterminée par la variance des risques inobservés au sein des DR et par la taille de la population des DR. Il existe des différences marquées entre la taille des populations des DR. Selon le recensement de 1976, les chiffres de la population des DR examinées dans l'étude allaient de 1 500 habitants à 2,1 millions d'habitants. Ainsi, une description de la répartition des risques nécessite l'utilisation simultanée des estimations ${}_nq(x)$, qui ne sont pas fiables au même degré.

L'étude vise à décrire d'une manière plus complète les conditions de mortalité au sein des populations nationales. A l'intérieur de chaque intervalle d'âge, on calcule des estimations non paramétriques des fonctions de densité de probabilité des risques de mortalité. Ces estimations fournissent une méthode souple d'évaluation de l'inégalité existant dans les

risques de mortalité (par exemple, au moyen d'indices de dispersion basés sur des quantiles).

Comme résultat de l'estimation de densité, on construit des tables de mortalité du moment en reliant les estimations des risques par âge qui correspondent à un quantile donné, et on les appelle tables de mortalité quantiles. Cette approche diffère du groupement plus conventionnel des estimations des risques par âge correspondant à une population définie géographiquement. On calcule les tables de mortalité quantiles pour des groupes d'âges (par exemple, en construisant des tables de mortalité quantiles abrégées). Ces dernières servent à illustrer une meilleure description de la mortalité pour une population nationale hétérogène. Dans d'autres buts que celui de donner un exemple, il serait préférable d'utiliser des tables de mortalité quantiles par année d'âge.

Le présent article insiste sur les tables de mortalité quantiles. Pour justifier la décision de les baser sur des données relatives aux DR, il faut tenir compte des différences de fiabilité entre les estimations $_{a}\hat{q}(x)$. C'est-à-dire que les tables de mortalité devraient refléter l'hétérogénéité dans la mortalité, plutôt que l'instabilité des estimations portant sur des petites régions. Il faut en outre que la variabilité entre les moyennes relatives aux DR soit aléatoire, pour que l'on puisse valablement décrire la répartition des risques par une densité de probabilité. Les deux sections qui suivent immédiatement cette introduction abordent ces aspects du problème d'estimation. La section 2 décrit l'estimateur $_{a}q(x)$ et une approximation de sa variance. Dans la section 3, on prétend que le

groupement géographique des données peut servir de base pour estimer la répartition des risques dans l'ensemble de la population canadienne.

L'estimation de la répartition des risques se fait en deux étapes: (1) la section 4 décrit l'ajustement des observations afin de déterminer la variabilité d'échantillonnage par une méthode linéaire empirique de Bayes, et (2) on obtient, à la section 5, des estimations non paramétriques ajustées de la densité de probabilité des risques, selon l'âge. La section 6 donne certains résultats explicatifs pour deux groupes d'âges, chez les hommes. La section 7 fournit des résultats tirés de certaines tables de mortalité quantiles. Enfin, la section 8 présente des remarques en guise de conclusion.

2. Variabilité d'échantillonnage

Le fait d'attribuer une variance d'"échantillonnage" aux estimations des quotients de mortalité peut être sujet à controverse si l'enregistrement des décès est complet et le sous-dénombrement du recensement négligeable. Dans ce cas, les taux de mortalité peuvent être considérés comme des valeurs exactes. Cette opinion ne tient aucun compte des facteurs ne se répétant pas et essentiellement imprévisibles qui influencent la mortalité d'une population pendant une période donnée (par exemple, températures extrêmes, etc.) Brillinger (1986) présente une base précise pour décrire la "variabilité naturelle", et applique un modèle à la mortalité des Canadiennes pendant la période 1926-1982. Divers autres modèles sont présentés dans les commentaires de discussion qui suivent l'étude de Brillinger.

On représente de la manière suivante les valeurs observées ${}_n\hat{q}(x)j$ pour la DR j selon l'équation de Chiang (1972), (en se servant de la notation actuarielle):

$$ {}_n\hat{q}(x)j = {}_nD(x)j / \{ [nP(x)j + (n-{}_na(x)) {}_nD(x)j] / n \} $$

$$ = {}_nD(x)j/K(x)j, \qquad (1) $$

dans laquelle ${}_nD(x)j$ correspond au nombre annuel moyen de décès pendant la période 1975-1977 chez les personnes âgées de x à x+n ans, ${}_nP(x)j$ représente la population du recensement de 1976, et ${}_na(x)$ désigne le nombre moyen d'années vécues dans l'intervalle d'âge par une personne décédée pendant la période observée. Le dénominateur qui figure dans l'équation 1 est une estimation de la population annuelle exposée au risque. Dans le reste de l'étude, on traitera le dénominateur K(x)j comme s'il s'agissait d'une valeur exacte.

Les nombres de décès par âge et DR sont tirés de totalisations spéciales fournies par la Division de la santé, de Statistique Canada. Les chiffres de la population proviennent de la base de données du recensement de 1976. Signalons que les chiffres de la population ont été arrondis de façon aléatoire à des multiples de 5, conformément aux exigences requises pour les totalisations du recensement destinées à l'usage public.

Pour les groupes d'âges (0-1 an) et (1-5) ans, ${}_na(x)$ était tiré des tables de mortalité masculine et féminine abrégées pour le Canada pendant la période 1975-1977 (Nagnur, 1986). Pour tous les autres groupes d'âges, on admettait en postulat que ${}_na(x)$ avait la valeur 2,5 .(c'est-à-dire, en supposant une répartition uniforme de décès). Ce dernier procédé représente une première approximation. On dispose d'autres approximations plus

complexes (voir, par exemple, Keyfitz et Frauenthal, 1975), mais il est difficile d'évaluer l'opportunité de ces méthodes lorsqu'on les applique à de petites populations qui sont soumises à des migrations importantes. Par conséquent, on n'utilise dans la présente étude qu'une première approximation, c'est-à-dire une stratégie préconisée par Hoem and Jensen (1982) et d'autres auteurs.

Désormais, on omettra, pour simplifier les expressions, le n souscrit indiquant l'étendue de l'intervalle d'âge ainsi que la notation représentant la dépendance à l'égard de l'âge. Toute l'analyse qui suit s'applique d'une manière indépendante à chaque intervalle d'âge.

On peut présenter un modèle statistique de la variabilité de $\hat{q}j$ de la manière suivante. Supposons que Qij soit une variable aléatoire de Bernoulli représentant des risques possibles de mortalité pour l'individu i, dans la DR j au sein de l'intervalle d'âge approprié. Qij prend la valeur 1 (qui représente un décès) avec la probabilité qij, et la valeur 0 avec la probabilité 1-qij. Ainsi, qij représente un risque personnel de décès.

Le nombre de décès hypothétiques dans la DR j est donné par la somme (supérieure à i) des risques personnels. Le nombre de décès observés (par exemple, Dj) correspond à la somme des réalisations $\hat{q}ij$ (prenant la valeur 1 en cas de décès de la personne et 0 dans les autres cas) de la variable aléatoire Qij, de sorte que l'équation 1 fournit un estimateur du risque moyen. La variance réelle du nombre de décès observés (en supposant des risques indépendants) est donnée par l'équation suivante:

$$\mathrm{Var}(Dj) = \sum qij\ (1 - qij), \qquad (2a)$$

que l'on peut récrire de la manière suivante:

$$Var(Dj) = Kj\ q.j\ (1-q.j) - \sum (qij - q.j)^2, \qquad (2b)$$

dans laquelle q.j représente la moyenne des risques personnels. L'équation 2b indique qu'une variance binomiale de Dj donne la variance maximale dans toutes les répartitions des risques personnels Kj avec la moyenne q.j. On peut améliorer l'exactitude de l'approximation binomiale en réduisant la largeur des intervalles d'âge, pourvu que Kj reste large. Le résultat dépend de la diminution de toutes les valeurs qij dans la limite. Signalons, toutefois, que l'hypothèse des risques indépendants est peut-être trop audacieuse.

Le reste de la présente étude traitera des étapes ultérieures d'analyse, en supposant que l'approximation binomiale soit adéquate. Ainsi, l'approximation de la variance d'échantillonnage de $\hat{q}j$ est $\hat{q}j(1-\hat{q}j)/Kj$. Il est à noter que les différences entre les valeurs Kj (par exemple, la taille de l'échantillon) entraînent des différences de fiabilité entre les valeurs $\hat{q}j$, et que l'on se base sur ce fait pour décrire la composante de la variance comme "variance d'échantillonnage".

3. Interprétation des différences de mortalité entre les DR

La construction de tables de mortalité régionales et la présentation de données sur la mortalité dans des atlas peuvent donner l'impression qu'il convient de traiter des populations distinctes géographiquement comme si elles étaient fermées aux migrations ou comme si, d'une manière ou d'une autre, les différences de mortalité dépendaient directement de la région. Toutefois, étant donné que les limites géographiques des DR sont déterminées

sans tenir compte des conditions de mortalité, les divergences entre les valeurs $\hat{q}j$ ne représentent pas des différences systématiques entre les populations des DR.

Même lorsque certains facteurs de risque de mortalité sont nettement liés à la géographie, il ne s'ensuit pas nécessairement que la géographie offre une base importante pour la stratification de la population exposée au risque (par exemple, en formant des sous-populations qui comportent des risques relativement homogènes). Dans la mesure où les facteurs de risque comprennent des habitudes ou modes de vie personnels, on peut s'attendre à ce qu'ils varient considérablement au sein des populations des DR.

A titre d'exemple, citons le cas du mélanome malin cutané. L'exposition répétée au soleil constitue un facteur de risque dans certains cas de cancer de la peau, et des facteurs purement géographiques comme l'altitude et le climat peuvent exercer une influence directe. Toutefois, dans le cas du mélanome, des influences prédominantes comprennent la sensibilité de l'individu et l'intermittence de l'exposition. Par conséquent, on pourrait représenter plus exactement des strates homogènes de risque en fonction de la profession (les personnes travaillant à l'intérieur étant exposées à un plus grand risque que celles qui travaillent à l'extérieur), de la sensibilité (peau blanche, tendance à brûler, etc.), et de la situation socio-économique (bronzage récréatif, vacances d'hiver, etc.), plutôt qu'en fonction seulement de la géographie (Elwood et autres, 1985).

Les caractéristiques des individus, dont l'appartenance à une population de DR plutôt qu'à une autre résulte d'un classement arbitraire, constituent une source de variabilité non systématique dans les estimations du risque moyen. Le classement peut correspondre à une répartition aléatoire de la population canadienne. Ainsi, la répartition des moyennes des DR reflétera la répartition des risques d'individus typiques au sein de la population nationale. C'est-à-dire qu'une densité de probabilité constitue une description adéquate de la répartition des moyennes des DR, étant donné que la source de variabilité comporte une interprétation des fréquences.

4. Ajustement linéaire empirique de Bayes destiné à corriger la variabilité d'échantillonnage

On peut utiliser les méthodes linéaires empiriques de Bayes (LEB) pour estimer des vecteurs de moyennes aléatoires. On peut représenter une forme simple de méthode LEB comme estimation d'un vecteur paramétrique aléatoire $\emptyset$ au moyen d'une régression sur un vecteur de données observées y (Robbins, 1983). L'objectif est de minimiser $E\{\emptyset(y))-\emptyset\}^a$, où $\emptyset(y)$ est une fonction linéaire d'y estimant $\emptyset$. Dans la présente application, les méthodes LEB consistent à tasser chaque qj vers le centre de la répartition. L'ampleur du rétrécissement est grande si la variance d'échantillonnage d'un $\hat{q}j$ donné est importante (c'est-à-dire en cas d'ajustement destiné à corriger la variabilité d'échantillonnage).

Le rétrécissement des estimations dans le sens de leur moyenne globale fournit un compromis raisonnable entre l'utilisation de la moyenne globale

pour représenter chaque DR (c'est-à-dire en supposant effectivement que la variance entre les DR est attribuable à la variabilité d'échantillonnage) ou l'acceptation de chaque estimation de DR sans modification (c'est-à-dire en supposant effectivement que les DR représentent des populations sans rapport entre elles).

On utilise une transformation de stabilisation de la variance afin que chaque $\hat{q}j$ ait une variance d'échantillonnage (asymptotique) indépendante de q.j. La stabilisation de la variance est donnée par la modification de la transformation de l'arc sinus d'Anscombe (1949):

$$\hat{t}j = 2 \text{ arc sinus } (\sqrt{(Dj + 3/8)/(Kj + 3/4)}). \qquad (3)$$

La modification d'Anscombe donne une stabilité accrue aux DR dans lesquelles K est petit, mais son effet est négligeable pour la plupart des DR examinées dans la présente application. Efron et Morris (1975) utilisent également une transformation de l'arc sinus appliquée à des données binomiales dans le cadre d'une méthode LEB.

La variabilité entre les moyennes des DR est représentée dans le modèle suivant:

$$\hat{t}j = t. + (tj - t.) + uj = t. + ej, \qquad (4)$$

où $tj = 2 \text{ arc sinus } (\sqrt{q.j})$.

Signalons que $Var(ej) = v + 1/Kj$, où $1/Kj$ représente la variance binomiale (d'échantillonnage) d'uj compte tenu de la transformation de l'arc sinus et v représente la variance qui dépasse la variance d'échantillonnage (c'est-à-dire la variance entre les écarts des DR (tj-t.)). Mentionnons en outre que $Cov(\hat{t}j, tj)=v$.

Les estimations de t et de v sont obtenues itérativement. L'estimation $\hat{t}.$ emploie l'inverse de la variance estimée de chaque tj comme coefficient de pondération:

$$\hat{t}. = \sum (\hat{t}j \; wmj) \; / \sum wmj, \qquad (5)$$

où wmj = $1/(\hat{v} + 1/Kj)$ avec sommation de m DR. Remarquons que si $\hat{v}$ était assez grand, les coefficients de pondération seraient presque égaux (c'est-à-dire à peu près indépendants de Kj).

Une estimation de v peut prendre la même forme qu'une estimation du maximum de vraisemblance en cas de normalité, avec un facteur de correction m/(m-1). On peut utiliser cette formule pratique de calcul:

$$\hat{v}i+1 = \lfloor m/(m-1) \rfloor \sum (\hat{t}j-\hat{t})^2 \; wvj \; / \sum wvj, \qquad (6)$$

où wvj = $\lceil 1/(1 + 1/ \hat{v}iKj) \rceil$, avec $\hat{v}i+1$ représentant une estimation actuelle et $\hat{v}i$ l'estimation antérieure. Il est à noter que $E(\hat{v}) = v$, que les écarts $(\hat{t}j-\hat{t})$ soient normaux ou autrement. Si l'on donne une valeur de départ à $\hat{t}$, on peut résoudre l'équation 6 d'une manière itérative. On estime alors $\hat{t}$ à partir de l'équation 5 avec de nouveaux coefficients de pondération, et on répète le cycle jusqu'à la convergence.

Le tableau 1 présente certains résultats des calculs de la variance. Il indique des coefficients de variation (CV) et des pourcentages de la variance qui dépassent la variance d'échantillonnage (appelés pourcentages d'excédent) pour chaque groupe d'âges. Il s'agit de mesures sommaires qui fournissent, d'une part, une comparaison des écarts-types non attribuables à l'échantillonnage par rapport à la moyenne et, d'autre part, la variance d'excédentaire relativement à la variance d'échantillonnage. Les CV oscillent entre 8 % et 28 %, ce qui indique que les écarts par rapport à la

moyenne sont, d'une manière typique,importants relativement à la moyenne. A l'exception du groupe de 0-1 an, les valeurs du pourcentage d'excédent sont toutes supérieures à 70 % (ce qui indique que la majeure partie de la variabilité présentée dans les données est une variabilité non attribuable à l'échantillonnage). Ces résultats montrent clairement qu'il existe des différences importantes entre les niveaux de mortalité présentés dans les données relatives aux DR. De même, la moyenne globale ne constitue pas un indicateur uniformément adéquat des risques courus dans la population.

On peut tirer des estimations LEB types de risque (q') de l'équation suivante:

$$t'j = 2 \text{ arc sinus } (\sqrt{q'j}) = \hat{t} + \left[\hat{v} / (\hat{v} + 1/Kj)\right] ej, \qquad (7)$$

dans laquelle le facteur de rétrécissement appliqué à ej estime le rapport entre $Cov(\hat{t}j, tj)$ et $/Var(\hat{t}j)$(c'est-à-dire d'une manière analogue à celle d'une pente d'une droite de régression). Si l'on poursuit l'analogie avec la régression, la variance résiduelle Var(t'j-tj) peut s'exprimer par la formule $Var(tj) - Cov(\hat{t}j,tj)^2/Var(tj)$, et est estimée par $\left[\hat{v}/(\hat{v} + 1/Kj)\right]$ (1/Kj).

Un certain nombre d'auteurs (notamment Louis, 1984, ainsi que Gaver et O'Muircheartaigh, 1987) se sont inquiétés de la possibilité que les estimateurs LEB types n'effectuent un rétrécissement trop poussé. Par exemple, l'observation selon laquelle la variance d'échantillon $\sum (t'j-\hat{t})^2/m-1$ est inférieure à $\hat{v}$ implique que la fonction de répartition empirique (FRE) de t'j n'approchera pas de la FRE de tj. Louis a proposé l'emploi d'un estimateur LEB restreint qui fournit une approximation de la FRE de tj, laquelle minimise l'écart par rapport aux estimations LEB types (c'est-à-dire en minimisant

$\sum wcj(t*j-t'j)^2)$:

$$t*j = 2 \text{ arc sinus } (\sqrt{q*j}) = [wcj/(wcj+d1)]\ t'j + [d2/(wcj+d1)], \qquad (8)$$

où wcj est une estimation de 1/Var(t'j-tj).

Les paramètres d_1 et d_2 sont des multiplicateurs de Lagrange choisis de façon à ce que la moyenne de l'échantillon $t*. = \sum t*j/m$ et la variance de l'échantillon $\sum (t*j-t*.)^2/m-1$ fassent pendant à $\hat{t}$ et $\hat{v}$, respectivement. Les valeurs de d_1 pour chaque sexe et groupe d'âges étaient obtenues au moyen d'un algorithme sécant. A la convergence, les valeurs approximatives de d_1 donnaient des écarts-types de l'échantillon qui différaient de $\hat{v}$ de moins de 0,00001 (c'est-à-dire de moins de 0,01 % d'erreur dans tous les cas). Après avoir déterminé les valeurs de d_1, on calculait le paramètre d_2 en réarrangeant l'équation 8 et on obtenait des moyennes de l'échantillon qui concordaient exactement avec $\hat{t}$ dans chaque cas.

Les estimations LEB restreintes t*j sont des estimations raisonnables des quantiles de risque dans la mesure où elles sont chacune proches de l'estimation LEB type correspondante et ont un rapport de concordance étroit avec la moyenne et la variance estimées de $\hat{t}j$. Ainsi, les estimations t*j servent de base à l'évaluation des fonctions de densité de probabilité du risque de mortalité après ajustement destiné à tenir compte des différences dans la fiabilité de $\hat{q}j$.

5. Estimations de la fonction de densité du risque

L'estimation non paramétrique d'une fonction de densité de probabilité du risque de mortalité vise à fournir une indication des caractéristiques

des répartitions dont il faudrait tenir compte dans un modèle paramétrique approprié. En ce sens, cette étape de l'analyse constitue une phase préliminaire. Toutefois, l'estimation de densité peut également donner des estimations de voisinage de pseudo-vraisemblance et des estimations d'échelle du risque de mortalité, et faciliter l'interpolation vers des quantiles de risque séparés par des espaces égaux.

On peut évaluer la qualité de l'ajustement d'un modèle de référence paramétrique simple de t* en considérant la linéarité ou l'absence de linéarité dans l'association entre les quantiles t* et les quantiles de référence prévus. Dans la présente application, les quantiles de référence (désignés ci-après par la lettre z) sont tirés d'une fonction de densité bêta. La densité bêta se présente sous la forme suivante:

$$f(z) = [z^{b-1} (1-z)^{c-1}] / \text{Bêta}(b-c), \qquad (9)$$

où Bêta (b,c) est la fonction bêta. Les densités bêta peuvent s'adapter à une grande variété de formes de répartition et se prêtent à la présentation de variables aléatoires limitées (t* est limitée par zéro et ∞). Les paramètres b et c sont choisis pour faire pendant à $\hat{t}/\pi$ et $\hat{v}/\pi^2$.

Silverman (1985) décrit un algorithme de lissage pour des fonctions spline cubiques, destiné à la régression non paramétrique, qui sera utilisé ici à la fois pour fournir des estimations ajustées de t* et des estimations ajustées des premières dérivées dt*/dz. L'algorithme minimise la fonction du vecteur:

$$(1-h)\,[t*-t*(z)]'W[t*-t*(z)] + h \int t*''(z)^2\, dt*(z), \qquad (10)$$

où t* est un vecteur des quantiles, t*(z) représente un vecteur des valeurs ajustées (à partir de la régression sur z), t*"(z) représente les secondes

dérivées des valeurs ajustées, et W est une matrice diagonale des coefficients de pondération. Les dérivées dt*(z)/dz estiment le rapport entre la probabilité (connue) de z et la probabilité de t*(z). Ainsi, on peut représenter la non-linéarité dans l'association entre t* et z par une densité empirique qui diffère de la densité bêta ajustée.

Le paramètre h figurant dans l'équation 10 est un paramètre de lissage, qui pondère par voie différentielle l'exactitude locale et l'ajustement global. A mesure que h s'approche de 1,0, l'algorithme représente la relation entre t* et z comme une transformation d'échelle de localisation (c'est-à-dire que la linéarité implique que t*"(z)=0). A mesure que h s'approche de 0,0, l'algorithme comprend implicitement autant de paramètres qu'il y a d'observations (c'est-à-dire à condition d'avoir une interpolation dans laquelle t*-t*(z)=0). On peut obtenir une valeur appropriée de h en minimisant les erreurs de contrevalidation (Silverman, 1984).

Comme les valeurs ajustées sont des quantiles estimés, on peut calculer W à partir de l'inverse des estimations des variances t*(z):

$$\mathrm{Var}\{t^*(z)\} = s^2 \left[(p(1-p)) \,/\, (m\, f\, t^*(z)^2) \right], \qquad (11)$$

où p est une estimation appropriée de la FRE, m représente le nombre d'observations, et s^2 est un scalaire (Parzen, 1979). On obtient les ordonnées de probabilité $f\{t^*(z)\}$ à partir des ordonnées de probabilité de z divisées par l'estimation correspondante de dt*(z)/dz. Ainsi, on peut calculer par itération l'algorithme de lissage pour des fonctions spline en mettant W chaque fois à jour grâce aux estimations améliorées de $f\{t^*(z)\}$ et s^2. La minimisation des erreurs de contrevalidation et l'usage itératif de

l'équation 11 ont permis d'obtenir des valeurs de h (pour chaque groupe d'âges), qui étaient situées dans l'intervalle de 0,76 à 0,99.

On trouve dans les études de Bofinger (1975) et Parzen (1979) des détails sur l'usage analogue de quantiles et de quantiles modèles en vue d'obtenir une estimation non paramétrique de densité. On justifie l'utilisation du lissage pour une fonction spline en considérant cette dernière comme une fonction de lissage (noyau) des moyennes mobiles, dotée d'une largeur de bande locale proportionnée à la variance locale (Silverman, 1984).

On se sert ensuite des estimations t*(z) pour construire des tables de mortalité. A cette fin, on calcule $\bar{q}^*$ en invertissant la transformation de l'arc sinus (c'est-à-dire $\bar{q}^*=[\sin(t^*(z)/2)]^2$) et on obtient des estimations de la fonction de densité de probabilité de $\bar{q}^*$ en appliquant le déterminant jacobien de la transformée inverse à f(t*(z)).

6. Résultats explicatifs

Cette section présente des résultats intermédiaires qui illustrent l'analyse. La figure 1 donne un tracé de $\hat{q}$ par rapport à $\bar{q}^*$ pour les hommes faisant partie des groupes de 0-1 an et 50-54 ans. On a choisi ces groupes d'âges parce que les répartitions étaient bien distinctes, mais pas au point de masquer complètement les détails (par exemple en réduisant la résolution du tracé). La figure 2 présente des tracés correspondants des ratios de densité empirique. Les ratios de densité, qui sont estimés à partir des premières dérivées ajustées obtenues ci-dessus, soulignent les différences

qui peuvent exister entre les fonctions de densité de risque de mortalité empirique et les transformations des fonctions de densité bêta ajustée. En ce sens, les ratios de densité facilitent l'évaluation de la qualité de l'ajustement. Signalons que les ratios de densité représentés graphiquement ont été ajustés pour tenir compte du facteur d'échelle constant π .

La pente d'une ligne résistante ajustant $\hat{q}j$ comme fonction de $\bar{q}*j$ peut fournir un indice général des effets du rétrécissement et du lissage. Chez les hommes de 0-1 an, la pente est de 1,33, ce qui indique que la dispersion des risques de mortalité non ajustés est généralement plus élevée dans une proportion de 33 % que celle des risques ajustés. La pente correspondante pour les hommes de 50-54 ans s'établit également à 1,08. Chez les hommes, la pente maximale est de 1,33 (groupe de 0-1 an), alors qu'elle est de 1,21 chez les femmes (de 35-39 ans). Les valeurs minimales correspondantes s'établissent à 1,03 chez les hommes de 20-24 ans et à 1,06 chez les femmes de 65-69 ans. L'effet des ajustements sur la dispersion des estimations des risques de mortalité n'est pas considérable; cette observation n'est pas compatible avec l'hypothèse selon laquelle les estimations de la mortalité dans les petites régions sont généralement instables.

La figure 1 illustre les effets du rétrécissement et du lissage sur la fonction $\hat{q}j$. Dans le groupe de 0-1 an, l'extrémité supérieure (y compris sept valeurs aberrantes possibles ou plus) a été considérablement réduite. Signalons que ce groupe d'âges à la plus faible proportion de variance non attribuable à l'échantillonnage de tous les groupes d'âges, ainsi que l'indiquent les valeurs du pourcentage d'excédent présentées au tableau 1.

Par conséquent, ce groupe d'âges est plus fortement influencé par le rétrécissement que les autres groupes d'âges, comme le montre la pente de la ligne résistante mentionnée plus haut.

La figure 2 donne une indication des différences qualitatives qui peuvent exister entre la répartition réelle des risques de mortalité et la densité bêta qui était choisie comme répartition de référence. Par exemple, les ratios de densité pour les hommes de 50-54 ans s'établissent, en général, à près de 1,0. Ainsi, pour ce groupe d'âges, la densité bêta peut fournir un modèle adéquat de probabilité. Peut-être est-il plus opportun d'utiliser un modèle à deux paramètres (c'est-à-dire la moyenne et la variance), qui semble convenir pour décrire la répartition des risques de mortalité pour ce groupe d'âges. Par contraste, les ratios de densité pour le groupe de 0-1 an représentent, chez les enfants en bas âge, des risques de mortalité qui sont moins obliques et ont des extrémités plus courtes que ce serait le cas dans une densité bêta. Cela est indiqué par la forme de la fonction du ratio de densité, c'est-à-dire par le sommet au centre (valeurs $> 1,0$) et les valeurs peu élevées (valeurs $< 1,0$) aux extrémités. On trouve plus souvent dans d'autres groupes d'âges des ratios de densité semblables à ceux des hommes de 0-1 an (c'est-à-dire qui indiquent un manque d'ajustement) que des ratios de densité qui ressemblent à ceux des hommes de 50-54 ans. Cela impliquerait qu'il serait peut-être nécessaire d'utiliser trois ou quatre paramètres pour évaluer la répartition des risques de mortalité pour la plupart des groupes d'âges.

Les résultats présentés ici sont influencés par les paramètres de rétrécissement qui figurent dans l'équation 8 (c'est-à-dire d_1 et d_2) ainsi

que par le choix du paramètre de lissage qui figure dans l'équation 10 (c'est-à-dire h). Ainsi, bien que les résultats puissent indiquer qu'un modèle à quatre paramètres serait plus approprié qu'un modèle à deux paramètres, ils ne constituent pas une base pouvant servir à des tests de signification (par exemple, des tests du rapport des vraisemblances).

7. Tables de mortalité quantiles

Les tables de mortalité du moment représentent les conditions de mortalité du point de vue des durées de vie synthétiques achevées (par exemple, en formant des courbes de survie à partir de l'intégrale du produit des probabilités de survie par âge $(1-\hat{q})$). Les tables de mortalité quantiles constituent une généralisation des tables de mortalité du moment construites sur des sections qui traversent les répartitions ajustées de mortalité par âge de $\bar{q}^*$. Ces sections relient les quantiles correspondants de divers groupes d'âges. Les courbes de survie construites peuvent servir à évaluer la variabilité pendant la durée de la vie (par exemple, l'espérance de vie à la naissance, ou l'âge médian au décès).

Le tableau 2 présente une comparaison de quotients de mortalité publiés antérieurement (Nagnur, 1986 - appelés estimations GLOBALES dans le tableau 2), avec des estimations qui utilisent les quotients de mortalité ajustés pour les DR et les fonctions de densité empirique. Cette comparaison vise à valider les résultats des tables de mortalité présentées ici en démontrant que, sur le plan de l'importance, les estimations sont comparables à des estimations plus conventionnelles. En outre, pour la plupart des groupes d'âges, l'ordre des valeurs MOYENNES, MÉDIANES et

MODALES indique clairement l'existence d'une asymétrie positive dans les répartitions de risque.

Toutefois, aucune des estimations présentées au tableau 2 n'est exactement comparable aux autres en tant qu'estimation de la même mesure de localisation. Plus précisément, les estimations GLOBALES et MOYENNES diffèrent tellement que les premières peuvent être présentées comme une moyenne pondérée de $\hat{q}j$ avec Kj comme coefficient de pondération, alors que les secondes se rapportent à une moyenne pondérée ayant pour coefficients de pondération les inverses des variances empiriques. L'utilisation d'une méthode de pondération semblable à celle employée dans l'équation 5 pourrait produire des différences nettement plus petites entre les coefficients de pondération par rapport aux différences existant entre les tailles de population. Ainsi, l'observation selon laquelle des estimations qui utilisent des données relatives aux DR sont souvent plus élevées que les estimations GLOBALES plus conventionnelles peut donner à entendre qu'il existe un biais dans les estimations GLOBALES. En raison des différences existant dans ces méthodes particulières d'estimation, la question de savoir quel est l'estimateur pouvant produire le plus grand biais est trop complexe pour qu'on puisse la résoudre ici. Néanmoins, des données de pondération selon la taille de la population risquent de produire des résultats inadéquats.

Le tableau 3 présente des estimations de la durée moyenne et médiane de vie tirées de certaines tables de mortalité quantiles. Ces estimations indiquent que la variabilité dans les risques courus au sein de la population canadienne est compatible avec une vaste gamme de durées de vie.

On peut obtenir une vérification partielle de ces résultats en comparant les estimations de l'extrémité inférieure avec un groupe de la population canadienne qui a notoirement connu une forte mortalité pendant la même période. En 1976, l'espérance de vie de la population des Indiens inscrits (représentant environ 1 % de la population canadienne) était de 60 ans pour les hommes et de 66 ans pour les femmes (Rowe et Norris, 1985). Les estimations basées sur les tables de mortalité quantiles sont moins extrêmes que les estimations relatives à la population des Indiens inscrits, ce qui peut indiquer que les estimations quantiles sont prudentes.

Le tableau 3 donne un aperçu des tendances récentes de la mortalité au Canada. Pendant les deux dernières décennies, l'espérance de vie au Canada s'est accrue d'au moins un an pendant chaque période de cinq ans. De telles augmentations ne sont pas importantes comparativement aux intervalles quantiles présentés au tableau 3. A ce propos, il est bon de se demander si ces tendances résultent d'améliorations générales de la santé ou d'améliorations qui sont particulières à une partie de la population, c'est-à-dire de se poser la question de savoir si les répartitions des risques relatives à des périodes antérieures seraient décalées vers le haut ou si leurs extrémités supérieures seraient allongées par rapport aux répartitions de 1976. La distinction influe sur nos prévisions des tendances futures de la mortalité.

Les tables de mortalité quantiles servent à concentrer l'attention sur deux éléments de la description des conditions de mortalité dans des populations hétérogènes. Il s'agit du besoin de tenir compte des différences entre individus du même âge, et de la nécessité de considérer

les différences qui peuvent exister dans le même groupe d'individus à divers âges. On peut construire des courbes de survie en déterminant à la fois une répartition des risques à chaque âge et en traçant une voie à travers la suite des répartitions de risques. Il y a lieu de croire que les fumeurs qui arrêtent de fumer suivent une voie différente de celle des fumeurs qui ne cessent pas de fumer. Les tables de mortalité quantiles représentent des parcours traversant les répartitions des risques, qui sont, en un sens, déterminés à la naissance.

8. Discussion

On peut résumer les principaux résultats de cette étude de la manière suivante:

(a) La variabilité entre les quotients de mortalité par âge, pour de petites régions, reflète en majeure partie la variabilité qui existe dans les risques de mortalité plutôt que l'instabilité des estimations relatives aux petites régions.

(b) Les différences entres les moyennes des DR sont suffisamment grandes pour avoir de l'importance sur les plans descriptif ou de l'élaboration des politiques, même après l'ajustement destiné à tenir compte de la variance d'échantillonnage et après le lissage. Par exemple, les tendances récentes de la mortalité nationale représentées par des changements dans l'espérance de vie peuvent refléter une contraction dans les extrémités des répartitions des risques plutôt que des décalages au centre.

(c) Les tables de mortalité quantiles extrêmes donnent une description directe de l'étendue des risques de mortalité dans une population hétérogène

en représentant les meilleures et les pires durées de vie synthétiques qui sont compatibles avec les données. Les tables de mortalité quantiles fournissent également des estimations de la dispersion près du centre des répartitions des risques de mortalité.

L'interprétation de ces résultats dépend de l'hypothèse selon laquelle les populations des DR peuvent constituer des échantillons représentatifs de personnes exposées à des risques indépendants. Cette opinion est très différente de celle qui suppose explicitement ou implicitement la présence d'effets particuliers à la région. Par ailleurs, c'est également une opinion qui s'accorde mieux avec des conditions de mortalité dans lesquelles la maladie contagieuse joue un rôle relativement secondaire alors que le mode de vie joue un rôle important.

Ces résultats pourraient inciter le chercheur à effectuer une analyse plus approfondie des données sur la mortalité fragmentées à un niveau de précision détaillé. Les études complémentaires pourraient se servir de variables de fragmentation qui ne peuvent être considérées comme non systématiques (comme l'état matrimonial), en plus d'utiliser un niveau de résolution géographique plus poussé . Mais, pour l'instant la répétition de ce travail sur plusieurs périodes pourrait constituer le meilleur complément de la présente analyse.

REMERCIEMENTS

Je remercie mes collègues Jane F. Gentleman et David Binder, dont les commentaires ont permis d'améliorer tant le contenu technique que la présentation de ce document. Je tiens également à remercier deux collaborateurs anonymes pour leurs commentaires.

TABLEAU 1

Mesures sommaires de la variabilité
dans les risques de mortalité de la population
des divisions de recensement
(L'arc sinus transforme la $\hat{t}j$ de $\hat{q}j$)

	HOMMES		FEMMES	
GROUPE D'ÂGES	Coefficient de variation	Pourcentage d'excédent	Coefficient de variation	Pourcentage d'excédent
0-1	14.7	50.1	16.3	50.1
1-4	28.4	82.1	26.2	74.6
5-9	24.8	77.8	27.5	76.1
10-14	24.7	79.7	27.3	76.0
15-19	23.0	91.3	25.0	81.5
20-24	25.4	92.8	27.3	81.0
25-29	21.4	86.7	26.0	78.7
30-34	20.5	84.0	24.0	76.9
35-39	19.1	82.7	20.2	73.6
40-44	16.4	82.7	16.8	72.3
45-49	13.7	83.5	17.6	80.7
50-54	11.5	84.2	14.3	78.5
55-59	11.3	87.6	13.2	81.9
60-64	9.6	87.8	12.1	82.7
65-69	8.7	86.7	12.5	84.9
70-74	9.3	89.5	9.7	80.6
75-79	9.0	88.2	10.1	75.7
80-84	8.2	84.6	11.1	88.0
85-89	9.2	85.0	9.1	77.7

Coefficient de variation = $100\sqrt{\hat{v}} / \hat{t}$

Pourcentage d'excédent - Variance qui dépasse la variance d'échantillonnage en pourcentage de la variance totale = $100\ (m\ \hat{v}) / \sum (\hat{v} + 1/Kj)$

$\hat{t}$ - moyenne estimée, $\hat{v}$ - variance non attribuable à l'échantillonnage
Kj - taille de la population, m - 259 divisions de recensement

Tableau 2

Comparaison entre des estimations de localisation
pour des quotients de mortalité q

ÂGE	GLOBALES	ESTIMATIONS MOYENNES	MÉDIANES	MODALES
HOMMES				
0-1	0.0141	0.0149	0.0146	0.0143
1-4	0.0032	0.0042	0.0039	0.0034
5-9	0.0023	0.0033	0.0032	0.0029
10-14	0.0021	0.0030	0.0028	0.0025
15-19	0.0075	0.0092	0.0087	0.0077
20-24	0.0093	0.0117	0.0111	0.0099
25-29	0.0074	0.0092	0.0088	0.0081
30-34	0.0078	0.0098	0.0095	0.0089
35-39	0.0107	0.0123	0.0118	0.0110
40-44	0.0164	0.0183	0.0178	0.0169
45-49	0.0273	0.0284	0.0279	0.0270
50-54	0.0439	0.0442	0.0437	0.0428
55-59	0.0693	0.0694	0.0687	0.0675
60-64	0.1053	0.1050	0.1054	0.1068
65-69	0.1557	0.1507	0.1499	0.1486
70-74	0.2281	0.2224	0.2219	0.2224
75-79	0.3234	0.3179	0.3165	0.3174
80-84	0.4508	0.4467	0.4453	0.4459
85-89	0.5989	0.6214	0.6207	0.6131
FEMMES				
0-1	0.0114	0.0123	0.0120	0.0115
1-4	0.0025	0.0035	0.0033	0.0032
5-9	0.0015	0.0024	0.0023	0.0020
10-14	0.0013	0.0020	0.0018	0.0016
15-19	0.0026	0.0036	0.0034	0.0030
20-24	0.0027	0.0036	0.0034	0.0030
25-29	0.0028	0.0036	0.0034	0.0028
30-34	0.0038	0.0047	0.0044	0.0039
35-39	0.0059	0.0072	0.0070	0.0067
40-44	0.0091	0.0103	0.0101	0.0097
45-49	0.0152	0.0151	0.0147	0.0144
50-54	0.0218	0.0227	0.0225	0.0220
55-59	0.0339	0.0342	0.0338	0.0330
60-64	0.0512	0.0513	0.0511	0.0515
65-69	0.0800	0.0795	0.0794	0.0804
70-74	0.1267	0.1282	0.1279	0.1287
75-79	0.2010	0.2037	0.2024	0.2010
80-84	0.3222	0.3318	0.3324	0.3337
85-89	0.4866	0.4965	0.4919	0.4817

ESTIMATIONS GLOBALES - telles que publiées dans l'étude de Nagnur (1986)- basées sur les rapports entre le nombre total de décès et la population totale.

ESTIMATIONS MOYENNES, MÉDIANES, MODALES - telles que déterminées par les estimations empiriques de Bayes des risques de mortalité et les densités de probabilité empirique.

Tableau 3

Estimations quantiles de la durée de vie

Quantile	Espérance de vie à la naissance		Âge médian au décès	
	HOMMES	FEMMES	HOMMES	FEMMES
1%	63.4	71.4	68.9	77.0
5%	65.5	73.4	70.5	78.4
10%	66.5	74.3	71.2	79.1
25%	68.2	75.7	72.5	80.2
50%	69.8	77.1	73.8	81.1
75%	71.4	78.5	75.2	82.1
90%	72.8	79.7	76.4	83.0
95%	73.7	80.5	77.1	83.6
99%	75.3	82.0	78.4	84.8
ESTIMATIONS MOYENNES	69.6	77.0	73.7	81.0
ESTIMATIONS MODALES	70.2	77.4	74.0	81.2
ESTIMATIONS GLOBALES	70.26	77.70	73.8	81.3

Les estimations MOYENNES et MODALES sont tirées des quotients de mortalité présentés au tableau 2.

ESTIMATIONS GLOBALES - l'espérance de vie présentée est celle qui est publiée dans l'étude de Nagnur (1986) - les âges médians au décès sont interpolés à partir de la courbe de survie publiée.

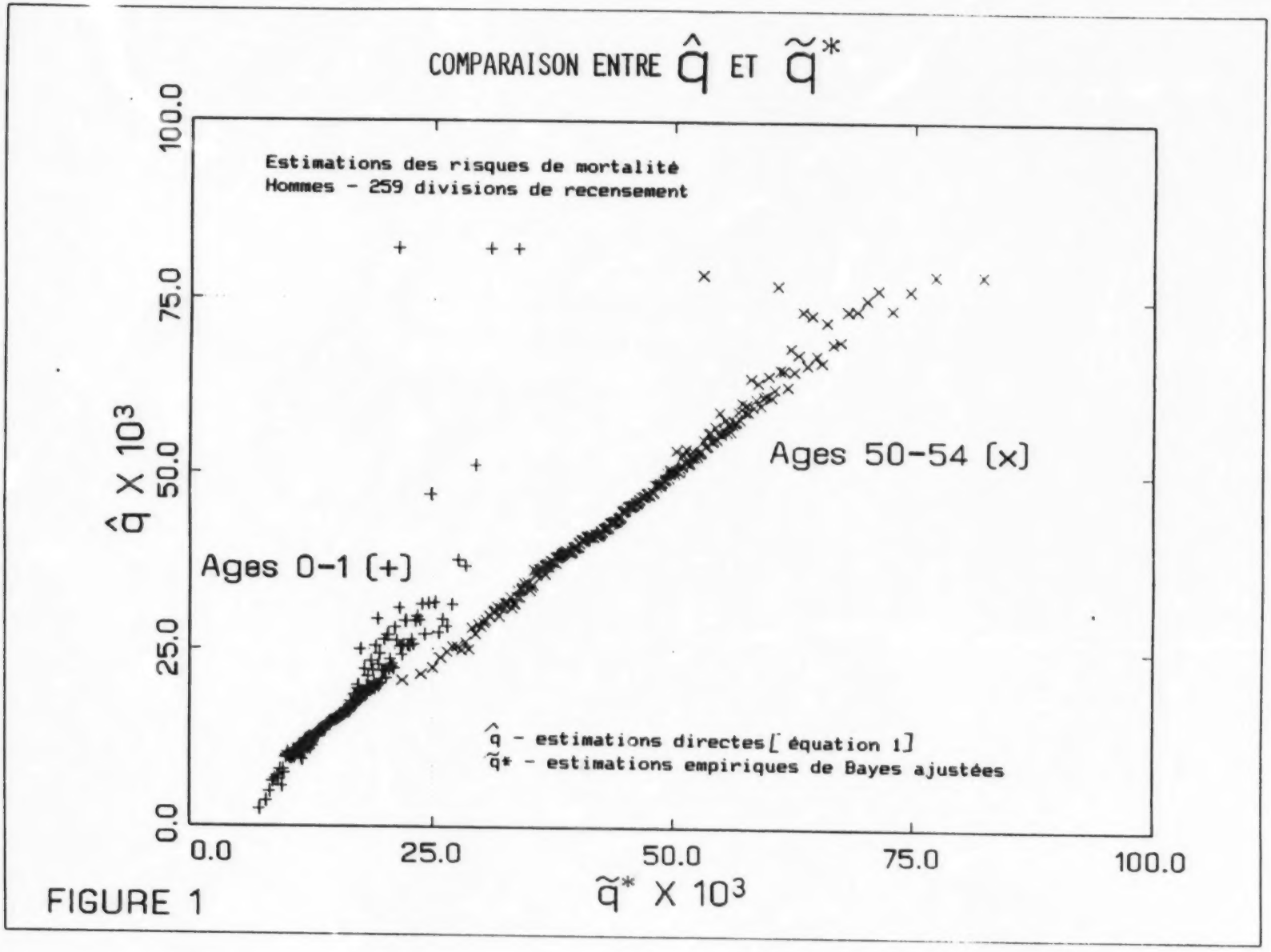

COMPARAISON ENTRE $\hat{q}$ ET $\tilde{q}^*$
Estimations des risques de mortalité
Hommes - 259 divisions de recensement
Ages 50-54 (x)
Ages 0-1 (+)
$\hat{q}$ - estimations directes[équation 1]
$\tilde{q}^*$ - estimations empiriques de Bayes ajustées
$\hat{q}$ X 10^3
$\tilde{q}^*$ X 10^3
0.0
25.0
50.0
75.0
100.0

FIGURE 1

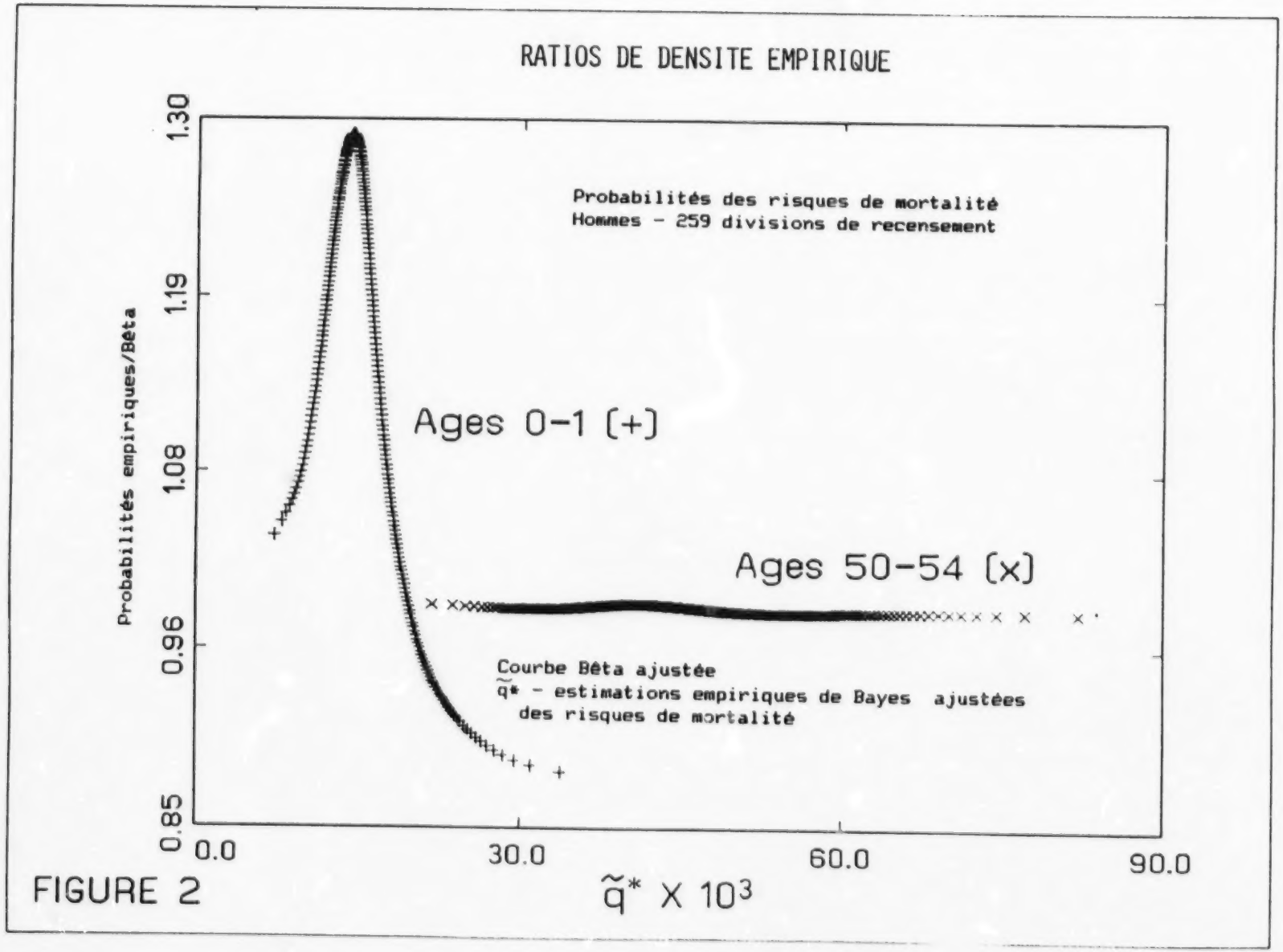

RATIOS DE DENSITE EMPIRIQUE
Probabilités des risques de mortalité
Hommes - 259 divisions de recensement
Ages 0-1 (+)
Ages 50-54 (x)
Courbe Bêta ajustée
q̃* - estimations empiriques de Bayes ajustées
des risques de mortalité
Probabilités empiriques/Bêta
1.30
1.19
1.08
0.96
0.85
0.0
30.0
60.0
90.0
q̃* X 10³

FIGURE 2

Bibliographie

Anscombe, F.J. (1948), "The transformation of poisson, binomial, and negative binomial data", Biometrika 35, 246-54.

Bofinger, E. (1975), "Estimation of a density function using order statistics", Austral. J. Statistics 17, 1-7.

Brillinger, David R. (1986), "The Natural Variability of Vital Rates and Associated Statistics", Biometrics 42, 693-734.

Chiang, C.L. (1972), "On constructing current life tables", JASA 67, 538-541.

Efron, Bradley et Morris, Carl (1975), "Data Analysis Using Stein's Estimator and Its Generalizations", JASA 70, 311-319.

Elwood, J.M., Gallagher, R.P., Hill, G.B., et Pearson, J.C.G., (1985), "Cutaneous Melanoma in Relation to Intermittent and Constant Sun Exposure - The Western Canada Melanoma Study", Int. J. Cancer 35, 427-433.

Gaver, Donald P., et O'Muircheartaigh, I.G. (1987), "Robust Empirical Bayes Analyses of Event Rates", Technometrics 29, 1-15.

Hoem, Jan M. et Jensen, Ulla Funck (1982), "Multistate Life Table Methodology: A Probabilist Critique", from Multidimensional Mathematical Demography, Kenneth C. Land et Andrei Rogers éditeurs, Academic Press, New York.

Keyfitz, Nathan et Frauenthal, James (1975), "An Improved Life Table Method", Biometrics 31, 889-899.

Louis, Thomas A. (1984), "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods", JASA 79, 393-398.

Nagnur, Dhruva (1986), Longévité et tables de mortalité chronologiques (abrégées) 1921-81, Canada et les provinces, n° 89-506 au catalogue, Statistique Canada, Ottawa.

Parzen, Emanuel, (1979), "Nonparametric Statistical Data Modeling", JASA 74, 105-131.

Robbins, Herbert, (1983), "Some Thoughts on Empirical Bayes Estimation", Annals of Statistics 11, 713-723.

Rowe, Geoff et Norris, M.J., (1985), "Projections de la mortalité des Indiens inscrits, 1982 à 1996", projet de rapport pour la Direction générale de la recherche, Orientation générale, Ministère des Affaires indiennes et du Nord Canada, Ottawa.

Silverman, B.W. (1984), "Spline Smoothing: The Equivalent Variable Kernel Method", Annals of Statistics 12, 898-916.

Silverman, B.W. (1985), "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting", J.R. Statis. Soc. B, 47, 1-52.

DIRECTION DES ÉTUDES ANALYTIQUES

DOCUMENTS DE RECHERCHE

No

1. RÉACTION COMPORTEMENTALE DANS LE CONTEXTE D'UNE SIMULATION MICRO-ANALYTIQUE SOCIO-ÉCONOMIQUE par Lars Osberg

2. CHÔMAGE ET FORMATION par Garnett Picot

3. DES PENSIONS AUX PERSONNES AU FOYER ET LEUR RÉPARTITION SUR LA DURÉE DU CYCLE DE VIE par Michael Wolfson

4. LA MODÉLISATION DES PROFILS D'EMPLOI DES CANADIENS AU COURS DE LEUR EXISTENCE par Garnett Picot

5. PERTE D'UN EMPLOI ET ADAPTATION AU MARCHÉ DU TRAVAIL DANS L'ÉCONOMIE CANADIENNE par Garnett Picot et Ted Wannell

6. SYSTÈME DE STATISTIQUES RELATIVES À LA SANTÉ: PROPOSITION D'UN NOUVEAU CADRE THÉORIQUE VISANT L'INTÉGRATION DE DONNÉES RELATIVES À LA SANTÉ par Michael Wolfson

7. PROJET PILOTE DE RACCORDEMENT MICRO-MACRO POUR LE SECTEUR DES MÉNAGES AU CANADA par Hans Adler et Michael Wolfson

8. NOTES SUR LES GROUPEMENTS DE SOCIÉTÉ ET L'IMPÔT SUR LE REVENU AU CANADA par Michael Wolfson

9. L'EXPANSION DE LA CLASSE MOYENNE: DONNÉES CANADIENNES SUR LE DÉBAT SUR LA DÉQUALIFICATION par John Myles

10. LA MONTÉE DES CONGLOMÉRATS par Jorge Niosi

11. ANALYSE ÉNERGÉTIQUE DU COMMERCE EXTÉRIEUR CANADIEN: 1971 et 1976 par K.E. Hamilton

12. TAUX NETS ET BRUTS DE CONCENTRATION DES TERRES par Ray D. Bollman et Philip Ehrensaft

13. TABLE DE MORTALITÉ EN L'ABSENCE D'UNE CAUSE POUR LE CANADA (1921 À 1981): Une méthode d'analyse de transition épidémiologique par Dhruva Nagnur et Michael Nagrodski

14. DISTRIBUTION DE LA FRÉQUENCE D'OCCURENCE DES SOUS-SÉQUENCES DE NUCLÉOTIDES, D'APRÈS LEUR CAPACITÉ DE CHEVAUCHEMENT par Jane F. Gentleman et Ronald C. Mullin

15. IMMIGRATION ET LE CARACTÈRE ETHNOLINGUISTIQUE DU CANADA ET DU QUÉBEC par Réjean Lachapelle

16. INTÉGRATION DE LA FERME AU MARCHÉ EXTÉRIEUR ET TRAVAIL HORS FERME DES MEMBRES DES MÉNAGES AGRICOLES par Ray D. Bollman et Pamela Smith

17. LES SALAIRES ET LES EMPLOIS AU COURS DES ANNÉES 1980: Évolution des salaires des jeunes et déclin de la classe moyenne par J. Myles, G. Picot et T. Wannell

18. PROFIL DES EXPLOITANTS AGRICOLES DOTÉS D'UN ORDINATEUR par Ray. D. Bollman

19. RÉPARTITIONS DES RISQUES DE MORTALITÉ: Une analyse de tables de mortalité par Geoff Rowe

Pour de plus amples renseignements, s'adresser au Président, Comité d'études des publications, Direction des études analytiques, Édifice R.H. Coats, 24ième étage, Statistique Canada, Parc Tunney's, Ottawa, Ontario K1A 0T6.